The Biometrics Section of the American Statistical Association

BIOMETRICS

Table of Contents

Number 3 September 1949 Volume 5

Material for *Biometrics* should be addressed to the Chairman of the Editorial Board, Institute of Statistics, North Carolina State College, Raleigh, N. C.; and material for Queries should go to "Queries", Statistical Laboratory, Iowa State College, Ames, Iowa, or to any member of the committee.

Officers
American Statistical Association

Officers
Biometrics Section

Editorial Board
Biometrics

Membership dues in the American Statistical Association are $8.00 a year, of which $4.00 is for a year's subscription to the quarterly *Journal*, $.75 is for a year's subscription to *The American Statistician*. Members of the Association may subscribe to *Biometrics* at a special rate of $2.00. Dues for associate members of the Biometrics Section are $3.00 a year, of which $2.00 is for a year's subscription to *Biometrics*. Single copies of *Biometrics* are $1.00 each and annual subscriptions are $3.00. Subscriptions and applications for membership should be sent to the American Statistical Association, 1603 K Street, N.W., Washington 6, D. C.

Entered as second-class matter, May 25, 1945, at the post office at Washington, D.C., under the Act of March 3, 1879. Biometrics is published four times a year—in March, June, September and December—by the American Statistical Association for its Biometrics Section. Editorial Office: 1603 K Street, N.W., Washington 6, D.C.

OPTIMUM ALLOCATION AND VARIANCE COMPONENTS IN NESTED SAMPLING WITH AN APPLICATION TO CHEMICAL ANALYSIS

SOPHIE MARCUSE

U. S. Naval Research Laboratory,
Washington, D. C.

INTRODUCTION

A SAMPLING TECHNIQUE frequently used in chemical and physical analyses for estimating the mean of a population is that of multiple random subsampling, called nested sampling by P. C. Mahalanobis.[1] For instance, when determining the moisture content of cheese, a food chemist might wish to select his samples randomly from different lots, and again from different cheeses of each lot, and finally make duplicate determinations on each cheese. A primary objective in the statistical design of such a sampling procedure is to minimize the cost of obtaining the sample estimate if the desired degree of precision is fixed, or conversely, to maximize the precision of the estimate obtained from a given amount of expenditure including personnel, time, and equipment. The question arises as to how the number of sampling units at each level should be determined to meet these optimum requirements assuming equal frequencies in the subclasses.

It is assumed in this paper that at each classification level, the cost is proportional to the number of units sampled at this level, and that the cost per sampling unit is known. Thus the total cost is a linear function of the numbers of sampling units at the various levels, with coefficients representing the (known) costs per sampling unit at these levels. On the other hand, the precision of the mean yielded by the experiment can be expressed in terms of the variance of this sample mean; it will then also be a linear function of the variances corresponding to each level, with coefficients involving the reciprocals of the number of units at the various levels. If the variances at the various levels are not known, they should be estimated from a preliminary experiment. The present paper discusses optimum allocation of the sampling units in nested sampling in terms of 3 levels. As an illustration of an experimental situation, a numerical example is given involving the estimation of variance components. In the appendix, the formulas for optimum allocation in nested sampling with k levels are derived.

[1] For reference see M. Ganguli's paper on Nested Sampling [7].

For concreteness, we consider the above mentioned specific problem of planning in the most economical way an experiment in food chemistry designed to determine the moisture content of cheese, the subsampling levels involving lots, cheeses, and determinations. Clearly, the principles elucidated in terms of this particular problem for 3 levels are applicable to a wider class of problems involving more levels in subsampling, as, for instance, by expanding this simplified experiment to more than one factory. Also, they may be applied to other than chemical investigations involving nested sampling, for instance: in the determination of the breaking strength of a certain type of bronze, a metallurgist may wish to choose random samples from different ladles, then again from different molds of each ladle, and make duplicate determinations on the samples from each mold; in a manufacturing process, the subsampling categories may be lots, bags, and batches; in a gunnery experiment, test shooting may be done by different operators taking a number of observations on different runs; in agricultural investigations, the entire area under survey may be subdivided into a large number of zones, these in turn into a large number of smaller zones, and so on; in studies of spray deposit in insect work, plots, trees, and apple samples have been used as subsampling levels [2]. Examples of nested sampling in biological and industrial work together with analyses of variance components may be found in G. W. Snedecor's [10] and L. H. C. Tippett's [12] books. In designing a sample survey for estimating the jute crop in India, P. C. Mahalanobis [9] has used the cost function for considerations of optimum allocation and discussed their general application to large scale sample surveys; principles of optimum allocation in nested sampling have been used by M. H. Hansen et al. [8] in a sample survey of business involving 2-fold nested sampling from finite populations (countries, stores), and by L. H. C. Tippett [12] who describes an experiment where in obtaining soil samples from counts of cysts, a number of "borings" of soil were taken and then several counts made on each boring.

DEFINITION OF NESTED SAMPLING

The problem considered is one in which the total population is subdivided into primary sampling units (lots); these in turn are subdivided into secondary sampling units (cheeses) on which several measurements (determinations) are made representing the tertiary sampling units. The nested sample is obtained by selecting at random first n_1 primary (lots), then n_2 secondary (cheeses), and finally n_3 tertiary sampling units (determinations) from each of the preceding units, where n_1 , n_2 ,

n_3 represent the class frequencies. A measure of the variance of the sample mean in terms of the class frequencies is desired. Before deriving it, the structure of the mathematical model will be explained.

Let x_{hij} denote the j-th determination from the i-th cheese of the h-th lot. Assuming that the effects of the sampling units at the different levels are additive, we may describe an individual observation x_{hij} in nested sampling [7] as:

$$x_{hij} = \mu + \xi_h + \eta_{hi} + \zeta_{hij} \tag{1}$$

$h = 1, 2, \cdots, n_1$ where h refers to the lot of cheese
$i = 1, 2, \cdots, n_2$ where i refers to the cheese in each lot
$j = 1, 2, \cdots, n_3$ where j refers to the determination on each cheese.

The value μ represents the general population mean and is thus a fixed constant. The components ξ_h, η_{hi}, ζ_{hij} are random variables with means and covariances equal to zero and with variances equal to σ_1^2, σ_2^2, σ_3^2, respectively, called variance components. Thus the components ξ_h, η_{hi}, ζ_{hij} represent the effects peculiar to the lots, cheeses, and determinations, and the variance components the variabilities at the different levels.

VARIANCE OF SAMPLE MEAN AND ESTIMATION OF VARIANCE COMPONENTS IN NESTED SAMPLING

From the definition of an individual observation x_{hij} in nested sampling, given by equation (1), we have for the sample mean

$$\bar{x} = \mu + \frac{\sum_{h=1}^{n_1} \xi_h}{n_1} + \frac{\sum_{h=1}^{n_1}\sum_{i=1}^{n_2} \eta_{hi}}{n_1 n_2} + \frac{\sum_{h=1}^{n_1}\sum_{i=1}^{n_2}\sum_{j=1}^{n_3} \zeta_{hij}}{n_1 n_2 n_3} \tag{2}$$

Then because of the assumptions made for the random variables ξ_h, η_{hi}, ζ_{hij} we obtain for the variance of the sample mean

$$\sigma_{\bar{x}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} + \frac{\sigma_3^2}{n_1 n_2 n_3} \tag{3}$$

This expression gives the variance or precision of the sample mean as a linear function of the reciprocals of n_1, $n_1 n_2$, and $n_1 n_2 n_3$ representing the total number of lots, cheeses, and determinations used. The coefficients are the variance components σ_1^2, σ_2^2, σ_3^2, being the variances encountered at the 3 subsampling levels.

As long as the parameter values σ_1^2, σ_2^2, σ_3^2 are unknown, the variance function $\sigma_{\bar{x}}^2$ in (3) cannot be used for solving the problem to determine the optimum values of the class frequencies. On the other hand, if a set of class frequencies were given and used in performing an experiment in nested sampling, then the unknown parameters σ_1^2, σ_2^2, σ_3^2 could be estimated from an analysis of variance of the experimental data. This dilemma[2] may be evaded by first carrying out a preliminary experiment in nested sampling[3] using a set of arbitrarily chosen class frequencies. We will show how the data obtained from such a preliminary experiment give advance estimates of σ_1^2 , σ_2^2 , σ_3^2 , say s_1^2 , s_2^2 , s_3^2 , to be used for estimating the coefficients of the variance function.

TABLE 1

ANALYSIS OF VARIANCE IN 3-FOLD NESTED SAMPLING

Source of Variation	Degrees of freedom	Mean Square	Expected Mean Square
Primary sampling units	$n_1^* - 1$	MS_1	$\sigma_3^2 + n_3^*\sigma_2^2 + n_3^*n_2^*\sigma_1^2$
Secondary sampling units within primary units	$n_1^*(n_2^* - 1)$	MS_2	$\sigma_3^2 + n_3^*\sigma_2^2$
Tertiary sampling units within secondary units	$n_1^*n_2^*(n_3^* - 1)$	MS_3	σ_3^2

Denote by n_1^*, n_2^*, n_3^* the given class frequencies of the preliminary experiment in nested sampling. Perform a customary analysis of variance on the observed data, as shown in the first 3 columns of table 1, where MS_1 , MS_2 , and MS_3 denote the mean squares corresponding to the primary, secondary, and tertiary sampling units. It can be shown that the expected values of the mean squares MS_1 , MS_2 , and MS_3 are the expressions shown in the last column of table 1[4]. Considering the estimates of these expressions by substituting the estimated variance components s_1^2 , s_2^2 , s_3^2 , we obtain the equations

$$\begin{aligned} MS_1 &= s_3^2 + n_3^*s_2^2 + n_3^*n_2^*s_1^2 \\ MS_2 &= s_3^2 + n_3^*s_2^2 \qquad (4) \\ MS_3 &= s_3^2 \end{aligned}$$

[2] See M. Friedman's discussion of a similar situation in planning an experiment ([11], p. 345).

[3] Or a mixed model design of experiment (e.g. randomized blocks or split plot) which includes the subsampling categories under consideration. Note that such a design might involve more degrees of freedom thus increasing the reliability of the estimated variance components ([3], [4]).

[4] Results for any number of sub-samplings and unequal frequencies are given by M. Ganguli [7].

Whence we have the solutions

$$\begin{aligned} s_3^2 &= MS_3 \\ s_2^2 &= \frac{MS_2 - MS_3}{n_3^*} \\ s_1^2 &= \frac{MS_1 - MS_2}{n_2^* n_3^*} \end{aligned} \qquad (5)$$

in which the estimated variance components are expressed in terms of the mean squares calculated in the analysis of variance table of the experimental data from nested sampling.[5] These equations can be extended from three to k subsamplings by the same reasoning.

OPTIMUM ALLOCATION IN 3-FOLD NESTED SAMPLING

The variance of the sample mean and the total cost expenditure for determining it, expressed in terms of the class frequencies, are the two functions needed for solving the optimum allocation problem under consideration. Considering the case of 3 levels, let $C(n_1, n_2, n_3)$ be the cost function and $V(n_1, n_2, n_3)$ the variance function, the variables n_1, n_2, n_3 representing the class frequencies. As given by equation (6), the cost function $C(n_1, n_2, n_3)$ is assumed to be an additive function of the costs at the three levels, that is the costs of n_1 primary, n_1n_2 secondary, and $n_1n_2n_3$ tertiary sampling units altogether, the cost per primary, secondary, and tertiary sampling unit being c_1, c_2, and c_3 respectively. The variance function $V(n_1, n_2, n_3)$ is given by equation (3) showing the variance of the sample mean, $\sigma_{\bar{x}}^2$, in 3-fold nested sampling; its parameters may be estimated from the data of a preliminary experiment by the analysis of variance procedure for estimating variance components as described above. Thus we have:

$$C(n_1, n_2, n_3) = c_1n_1 + c_2n_1n_2 + c_3n_1n_2n_3 \qquad (6)$$

$$V(n_1, n_2, n_3) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1n_2} + \frac{\sigma_3^2}{n_1n_2n_3} \qquad (3)$$

The problem of optimum allocation is to minimize $C(n_1, n_2, n_3)$ by proper choice of n_1, n_2, n_3 subject to the constraint that the allowable

[5] This analysis of the variance components was performed on data from nested sampling, which is a special case of Model II analysis of variance as shown below. If a similar analysis of variance components is routinely carried out on data belonging to Model I, the interpretation differs. In Model II, the computed variance components estimate the variances σ_1^2, σ_2^2, σ_3^2 associated with random factors, whereas in Model I, these are dummy symbols representing sums of squares of differences related to the variation of systematic (or fixed) factors ([1], [5]).

amount of variance is preassigned, say v, or to minimize $V(n_1, n_2, n_3)$ by proper choice of n_1, n_2, n_3 subject to the constraint that the total amount of cost is fixed, say c. Let n_{C1}, n_{C2}, n_{C3} and n_{V1}, n_{V2}, n_{V3} be the optimum solutions of the two problems respectively. By applying Lagrange multipliers it can be shown[6] that these optimum values of n_1, n_2, n_3 are

$$\begin{aligned} n_{C1} &= \frac{\sigma_1}{v} \frac{\sum_{i=1}^{3} (\sigma_i \sqrt{c_i})}{\sqrt{c_1}} \\ n_{C2} &= \frac{\sigma_2}{\sigma_1} \sqrt{\frac{c_1}{c_2}} \\ n_{C3} &= \frac{\sigma_3}{\sigma_2} \sqrt{\frac{c_2}{c_3}} \end{aligned} \tag{7}$$

$$\begin{aligned} n_{V1} &= \frac{\sigma_1}{\sum_{i=1}^{3} (\sigma_i \sqrt{c_i})} \frac{c}{\sqrt{c_1}} \\ n_{V2} &= \frac{\sigma_2}{\sigma_1} \sqrt{\frac{c_1}{c_2}} \\ n_{V3} &= \frac{\sigma_3}{\sigma_2} \sqrt{\frac{c_2}{c_3}} \end{aligned} \tag{8}$$

The sets of equations (7) and (8) show similar features. Except for the first level, the optimum combination of the number of sampling units is independent of the given degree of precision or the fixed total cost, being the same whether the precision or the amount of cost is assigned beforehand. Therefore, when planning an experiment in nested sampling the analyst need be concerned with the given cost or precision only in selecting the number of primary sampling units. Clearly, an increase in funds would be utilized most efficiently, that is resulting in the highest possible precision, by a proportional increase in the number of primary sampling units, and similarly, the most economical way for attaining a higher degree of precision would consist in choosing a correspondingly greater number of primary sampling units.

In many instances, the research analyst might not wish to depend

[6] See appendix for development of these formulas.

on considerations of optimum allocation in the choice of the frequencies at all levels, but might prefer to take, for instance, duplicate or triplicate determinations from each cheese for check purposes, thus preassigning the class frequency associated to the tertiary sampling unit, n_3 . If n_3 is prefixed, the corresponding optimum allocation formulas[7] are

$$\begin{aligned} n'_{C1} &= \frac{\sigma_1}{v} \frac{\left[\sigma_1 \sqrt{c_1} + \sqrt{\left(\sigma_2^2 + \frac{\sigma_3^2}{n_3}\right)(c_2 + c_3 n_3)}\right]}{\sqrt{c_1}} \\ n'_{C2} &= \frac{\sqrt{\sigma_2^2 + \frac{\sigma_3^2}{n_3}}}{\sigma_1} \sqrt{\frac{c_1}{c_2 + c_3 n_3}} \end{aligned} \qquad (9)$$

in the case that the variance v is given; and

$$\begin{aligned} n'_{V1} &= \frac{\sigma_1}{\left[\sigma_1 \sqrt{c_1} + \sqrt{\left(\sigma_2^2 + \frac{\sigma_3^2}{n_3}\right)(c_2 + c_3 n_3)}\right]} \frac{c}{\sqrt{c_1}} \\ n'_{V2} &= \frac{\sqrt{\sigma_2^2 + \frac{\sigma_3^2}{n_3}}}{\sigma_1} \sqrt{\frac{c_1}{c_2 + c_3 n_3}} \end{aligned} \qquad (10)$$

in the case that the total cost c is given.

NUMERICAL EXAMPLE

The figures shown in table 2 are results from analyses of samples of cheese for the determination of moisture content.[8] They will serve as the preliminary data for obtaining estimates of the variance components. The experimental set-up in nested sampling involves duplicate determinations made on 2 cheeses from each of 3 lots, the different cheeses and the different lots being randomly selected ($n_1^* = 3$, $n_2^* = 2$, $n_3^* = 2$).

The first 4 columns of table 3 show the results of an analysis of variance of these data. In nested sampling the sums of squares may be calculated as follows: Consider first table 2 (in which there are 3 factors: duplicates, cheeses, and lots) and refer to the figures, representing 1 determination, as "totals." Subsequently, obtain the totals

[7] See appendix for development of formulas in which all but the first k' are fixed.

[8] The data are drawn from "Report on Sampling Fat and Moisture in Cheese" by William Horwitz and Lila F. Knudsen, J. Ass. Off. Agr. Chem., vol. 31 (1948), pp. 300–306; slight modifications have been made for illustrative purposes. The author acknowledges the suggestions of Lila F. Knudsen;

TABLE 2

MOISTURE CONTENT OF 2 CHEESES FROM EACH OF 3 DIFFERENT LOTS, DETERMINED 2 TIMES

Cheese	Lot		
	I	II	III
1	39.02 38.79	35.74 35.41	37.02 36.00
2	38.96 39.01	35.58 35.52	35.70 36.04

for the duplicates on each cheese (there remain 2 factors: cheeses and lots), and also the totals of the 4 determinations on each lot (there remains 1 factor: lots), in addition to the total for the entire table (no factor remains). Denote by Q_3 , Q_2 , Q_1 , and Q_0 the sum of squares of these corresponding totals divided by the number of determinations making up each total:

TABLE 3

ANALYSIS OF VARIANCE OF DATA ON MOISTURE CONTENT OF CHEESE GIVEN IN TABLE 2

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square	Estimated Variance Components
Lots	2	$SS_1 = 25.9001$	$MS_1 = 12.9501$	$\sigma_3^2 + 2\sigma_2^2 + 4\sigma_1^2$	$s_1^2 = 3.2028$
Cheeses within lots	3	$SS_2 = .4166$	$MS_2 = .1389$	$\sigma_3^2 + 2\sigma_2^2$	$s_2^2 = .0143$
Determinations within cheeses	6	$SS_3 = .6620$	$MS_3 = .1103$	σ_3^2	$s_3^2 = .1103$

$$Q_3 = 39.02^2 + 38.79^2 + \cdots + 35.70^2 + 36.04^2 = 16{,}365.5607$$

$$Q_2 = \frac{77.81^2 + 77.97^2 + 71.15^2 + 71.10^2 + 73.02^2 + 71.74^2}{2}$$

$$= 16{,}364.8988$$

$$Q_1 = \frac{155.78^2 + 142.25^2 + 144.76^2}{4} = 16{,}364.4821$$

$$Q_0 = \frac{442.79^2}{12} = 16{,}338.5820$$

Then the sums of squares in analysis of variance, SS_1, SS_2, SS_3, are the successive differences of these expressions:

$$SS_1 = Q_1 - Q_0 = 25.9001$$

$$SS_2 = Q_2 - Q_1 = 0.4166$$

$$SS_3 = Q_3 - Q_2 = 0.6620$$[9]

The sums of squares and the corresponding mean squares are shown in columns 3 and 4 of table 3. The estimated variance components s_1^2, s_2^2, s_3^2, shown in the last column of table 3, follow from equations (5). These values represent the advance estimates from the preliminary data to be used in the planning of the experiment.

The problem of designing an experiment with optimum allocation may arise in chemical laboratory work, e.g., when it is desired to set up in the most economical way routine analyses of samples of cheese for the determination of moisture content. In the example under consideration we assume that the chemist wants to spend not more than 60 dollars altogether to be allocated in such a way that the highest precision results; that he requires duplicate determinations for check purposes; and that the cost factors per lot, cheese, and determination are 10, 3, and 1 dollar respectively. Since these requirements prefix the class frequency n_3 and the total cost C, formulas (10) are appropriate. Substituting $n_3 = 2$, $c = 60$, $c_1 = 10$, $c_2 = 3$, and $c_3 = 1$, and for the variances σ_1^2, σ_2^2, σ_3^2 their estimates $s_1^2 = 3.2028$, $s_2^2 = 0.0143$, $s_3^2 = 0.1103$, we obtain:

$$n'_{V1} = 5.43 \qquad n'_{V2} = 0.21$$

The corresponding integer values have to be chosen in accordance with the conditions of the experiment. Since n_2, the number of cheeses selected from each lot, must be at least one, the number of lots, n_1, may be reduced. An examination of the integers smaller than n'_{V1} shows that $n_1 = 4$ together with $n_2 = 1$ fulfill the required conditions. Thus 4 lots and 1 cheese give the optimum solution for the problem under consideration.

The merit of this optimum combination may be judged by comparing it to other combinations of class frequencies. In table 4 a number of various combinations (columns 1 and 2) are presented together with the precision of the sample mean (columns 5 and 6) and

[9] Using the figures given for Q_2, Q_3 above, we have $Q_3 - Q_2 = .6619$ instead of .6620. Such a difference in the last decimal place is due to rounding off results, intermediate computations being carried out to more decimal places.

TABLE 4

ESTIMATED PRECISION AND COST OF DETERMINING MOISTURE CONTENT OF CHEESE WHEN A SPECIFIED NUMBER OF LOTS (n_1) AND A SPECIFIED NUMBER OF CHEESES FROM EACH LOT (n_2) ARE USED AND TWO DETERMINATIONS ($n_3 = 2$) ARE MADE ON EACH CHEESE. CONSTANTS USED ARE ADVANCE ESTIMATES CALCULATED FROM PRELIMINARY DATA (TABLES 2 AND 3).

Formulas used:	Constants used:
$N = n_1 n_2 n_3$	$n_3 = 2$
$C = c_1 n_1 + c_2 n_1 n_2 + c_3 n_1 n_2 n_3$	$c_1 = 10, c_2 = 3, c_3 = 1$
$V = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} + \frac{s_3^2}{n_1 n_2 n_3}$	$s_1^2 = 3.2028, s_2^2 = .0143, s_3^2 = .1103$
$CV = \frac{\sqrt{v}}{\bar{x}} \times 100$	$\bar{x} = 36.90$

Number of—		Expenditure		Estimated Precision	
Lots	Cheeses	Number of Determinations	Total Cost in dollars	Variance of mean	Coefficient of Variation
n_1	n_2	N	C	V	CV
(1)	(2)	(3)	(4)	(5)	(6)
5	3	30	125	0.6452	2.18
5	2	20	100	0.6475	2.18
5	1	10	75	0.6544	2.19
4	3	24	100	0.8065	2.43
4	2	16	80	0.8094	2.44
4	1	8	60	0.8181	2.45
3	3	18	75	1.0753	2.81
3	2	12	60	1.0792	2.82
3	1	6	45	1.0907	2.83
2	3	12	50	1.6130	3.44
2	2	8	40	1.6188	3.45
2	1	4	30	1.6361	3.47
1	3	6	25	3.2259	4.87
1	2	4	20	3.2375	4.88
1	1	2	15	3.2722	4.90

the expenditure involved in determining it (columns 3 and 4). Column 3 shows the total number of determinations made, the total cost is given in column 4, and column 6 compares the relative precision of the sample mean, indicated by its coefficient of variation, to the absolute precision in terms of the variance (column 5). Duplicate determinations are used throughout. It can be seen that the 4-1-2 combination is more economical than the 3-2-2 combination—the one used in the preliminary experiment—since it obtains a higher precision but requires the same cost (60 dollars). Also, the combination 3-2-2 is less efficient than the combination 3-1-2 since, for the same precision, the latter combination needs half the number of determinations and requires only 45 dollars instead of 60 dollars. In general, it pays to increase the number of lots instead of the number of cheeses since the former are more variable.

REMARKS ON NESTED SAMPLING AS A SPECIAL CASE OF MODEL II ANALYSIS OF VARIANCE

The mathematical model of nested sampling as given by the fundamental equation (1) and its assumptions, is closely related to one specific mathematical model used in analysis of variance. Two models of analysis of variance, usually referred to as Model I and Model II, have been discussed recently by S. L. Crump [3] and C. Eisenhart [5]. It seems worthwhile to show that, in virtue of the underlying assumptions, nested sampling represents a special case of Model II of analysis of variance.

The two different models of analysis of variance involve the analysis of two different types of factors: systematic factors in Model I and random factors in Model II. A factor such as "treatment" or "lot" is a random or a systematic factor depending on the way its variants are chosen. Here the term "variant" of a factor is used based on Fisher's terminology [6], for instance, the variants of the factor "treatment" may be e.g. "nitrogen" and "phosphate" and different lots the variants of the factor "lot." When an experimenter selects the two treatments "nitrogen" and "phosphate," he selects them systematically from a population of possible treatments on the basis of subject matter judgment; on the other hand, when selecting different lots of material for studying the effects of the treatments, he generally bases his choice on random selection ([5], [10] Chapter 8). Since systematically chosen variants produce systematic variation and randomly chosen variants random variation, the type of factor may be determined according to the issue: systematic or random variation. Usually, "methods" and "treatments" represent systematic factors, "blocks" and "lots" random factors, whereas factors such as "days" or "animals" or "locations" may represent either systematic or random factors; both types of factor will often occur in the same experiment; then the model is a mixed one.

Now the factors encountered in nested sampling are the primary, secondary, tertiary sampling units (lots, cheeses, determinations). Under the assumptions made, the variants of these factors, i.e. the units selected at each level, were chosen randomly. These factors, therefore, are random factors and thus nested sampling belongs to Model II.

In order to describe more accurately the relationship of nested sampling to Model II of analysis of variance, we subdivide the random factors of Model II into two categories: cross classified[10] with respect to another factor or not. For instance, in the 2 factor "day-animal" experiment discussed by C. Eisenhart [5] as an example of Model II, the random factor "animal" is cross classified with respect to the factor "days," each of the randomly chosen animals being tested on all days (the analysis of variance table contains: "Between days", "Between animals," and "Residual" with $d - 1$, and $a - 1$, and $(a - 1)(d - 1)$ degrees of freedom respectively). On the other hand, there would be no cross classification, if on each day a number of animals were randomly chosen for testing, as for instance in an inoculation experiment affecting the sensitivity of the animal (the analysis of variance contains: "Between days," and "Between animals within days" with $d - 1$, and $d(a - 1)$ degrees of freedom respectively). Likewise, no cross classification would be involved for the random factor "animal" if each animal would be tested on a couple of days which were randomly selected, as e.g. if only one animal could be tested per day (the analysis of variance contains: "Between animals," and "Between days within animals" with $(a - 1)$, and $a(d - 1)$ degrees of freedom respectively). Nested sampling represents the second category of Model II in which the random factors involved are not cross classified since for each primary sampling unit a number of secondary sampling units is selected randomly, and so on. The question as to which order of subsampling should be adopted in the nested sampling procedure, as, for instance, whether to use "animals" as primary sampling units and "days" as secondary sampling units, or conversely, is a decision to be made on the basis of subject matter judgment.

APPENDIX

We shall now derive the optimum values of the class frequencies, given for the three-fold level by formulas (7), (8), (9), and (10), for the general case of k-fold nested sampling. Instead of solving the prob-

[10] This term is not synonymous with "ordered". Note that items in table 2 below are ordered for purely designative reasons there being neither a cross classification nor an element of "sequence" involved.

lem directly by introducing the Lagrange multiplier, we will apply this procedure to a pair of generalized functions. We then obtain as special cases the solution formulas for optimum allocation in

i. k-fold nested sampling

ii. k-fold nested sampling in which some class frequencies are fixed beforehand

iii. stratified sampling from finite population (k strata, 2 levels).

a. *Minimum Problem for 2 Generalized Functions*

Let the two generalized functions be

$$F_1(N_1, \cdots, N_k) = \sum_{i=1}^{k} a_{1i}N_i + a_1 \tag{11}$$

$$F_2(N_1, \cdots, N_k) = \sum_{i=1}^{k} \frac{a_{2i}}{N_i} + a_2 \tag{12}$$

where $N_1, \cdots, N_k$ denote variables and a_1, a_2, and a_{1i}, a_{2i} ($i = 1, \cdots, k$) are constants.

Consider first the problem to minimize $F_1(N_1, \cdots, N_k)$ subject to the side condition

$$F_2(N_1, \cdots, N_k) = b_2 \tag{13}$$

where b_2 is a constant. Using the Lagrange multiplier λ in the usual way, we let the derivatives of $F_1 + \lambda F_2$ with respect to N_i ($i = 1, \cdots, k$) be zero, and obtain

$$a_{1i} - (\lambda a_{2i}/N_i^2) = 0$$

or

$$N_i = \sqrt{\lambda}\sqrt{a_{2i}/a_{1i}}$$

Substituting these values of N_i in (13), where F_2 is given by (12), we have

$$F_2(N_1, \cdots, N_k) = (1/\sqrt{\lambda}) \sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}} + a_2 = b_2$$

Therefore

$$\sqrt{\lambda} = \frac{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}}{b_2 - a_2}$$

Hence we obtain the optimum values

$$N_{1i} = \frac{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}}{b_2 - a_2} \sqrt{\frac{a_{2i}}{a_{1i}}} \tag{14}$$

Similarly, we obtain the solution of the problem to minimize $F_2(N_1, \cdots, N_k)$ subject to the side condition

$$F_1(N_1, \cdots, N_k) = b_1 \tag{15}$$

where b_1 is a constant:

$$N_{2i} = \frac{b_1 - a_1}{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}} \sqrt{\frac{a_{2i}}{a_{1i}}} \tag{16}$$

Now introduce the variables

$$n_1 = N_1, \qquad n_i = N_i/N_{i-1} \qquad (i = 2, \cdots, k) \tag{17}$$

then $N_i = n_1 \cdots n_i (i = 1, \cdots, k)$. Substituting the new variables in (11) and (12), we obtain the functions

$$f_1(n_1, \cdots, n_k) = \sum_{i=1}^{k} a_{1i} n_1 \cdots n_i + a_1 \tag{18}$$

$$f_2(n_1, \cdots, n_k) = \sum_{i=1}^{k} \frac{a_{2i}}{n_1 \cdots n_i} + a_2 \tag{19}$$

Substituting (14) in (17), we find that the minimum solutions of $f_1(n_1, \cdots, n_k)$ under the side condition $f_2(n_1, \cdots, n_k) = b_2$ are:

$$n_{11} = \frac{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}}{b_2 - a_2} \sqrt{\frac{a_{21}}{a_{11}}}$$

and

$$n_{1i} = \sqrt{\frac{a_{2i}a_{1,i-1}}{a_{1i}a_{2,i-1}}} \qquad (i = 2, \cdots, k) \tag{20}$$

Similarly, substituting (16) in (17), we find the minimum solutions of $f_2(n_1, \cdots, n_k)$ under the side condition $f_1(n_1, \cdots, n_k) = b_1$:

$$n_{21} = \frac{b_1 - a_1}{\sum_{i=1}^{k} \sqrt{a_{1i}a_{2i}}} \sqrt{\frac{a_{21}}{a_{11}}}$$

and

$$n_{2i} = \sqrt{\frac{a_{2i}a_{1,i-1}}{a_{1i}a_{2,i-1}}} \qquad (i = 2, \cdots, k) \tag{21}$$

Note that $n_{1i} = n_{2i} (i = 2, \cdots, k)$.

b. *Application to Optimum Allocation Problems in Sampling*

i. *Nested Sampling*

Substituting $a_{1i} = c_i$, $a_{2i} = \sigma_i^2$ and $a_1 = a_2 = 0$ in (18) and (19), we obtain the 2 functions

$$g_1(n_1, \cdots, n_k) = \sum_{i=1}^{k} c_i n_1 \cdots n_i \tag{22}$$

$$g_2(n_1, \cdots, n_k) = \sum_{i=1}^{k} \frac{\sigma_i^2}{n_1 \cdots n_i} \tag{23}$$

These functions represent the general case of the cost function $C(n_1, n_2, n_3)$ and the variance function $V(n_1, n_2, n_3)$ used above in section 4. Setting $b_1 = c$ and $b_2 = v$ yields the corresponding side conditions. Therefore applying formulas (20) and (21), we have as the minimum solutions of $g_1(n_1, \cdots, n_k)$ under the side condition $g_2(n_1, \cdots, n_k) = v$

$$n_{11} = \frac{\sigma_1}{v} \frac{\sum_{i=1}^{k} (\sigma_i \sqrt{c_i})}{\sqrt{c_1}}$$

and (24)

$$n_{1i} = \frac{\sigma_i}{\sigma_{i-1}} \sqrt{\frac{c_{i-1}}{c_i}} \qquad (i = 2, \cdots, k)$$

and as the minimum solutions of $g_2(n_1, \cdots, n_k)$ under the side condition $g_1(n_1, \cdots, n_k) = c$

$$n_{21} = \frac{\sigma_1}{\sum_{i=1}^{k} (\sigma_i \sqrt{c_i})} \frac{c}{\sqrt{c_1}}$$

and (25)

$$n_{2i} = \frac{\sigma_i}{\sigma_{i-1}} \sqrt{\frac{c_{i-1}}{c_i}} \qquad (i = 2, \cdots, k)$$

Specializing equations (24) and (25) to the case $k = 3$ yields equations (7) and (8). Specializing equation (25) to the case $k = 2$ and letting cost be expressed in terms of time, $c_1 = kt$, $c_2 = t$, gives equation 10.32 in L. H. C. Tippett's book [12].

ii. *Nested Sampling with Some Prefixed Class Frequencies*

Let $n_1', \cdots, n_{k'}'$ be the unknown frequencies and $n_{k'+1}, \cdots, n_k$ be fixed beforehand. The equations (22) and (23) may then be rewritten in terms of $n_1', \cdots, n_{k'}'$ as follows:

$$h_1(n_1', \cdots, n_{k'}') = \sum_{j=1}^{k'} c_j n_1' \cdots n_j' + n_1' \cdots n_{k'}' \sum_{l=1}^{k-k'} c_{k'+l} n_{k'+1} \cdots n_{k'+l} \tag{26}$$

$$= \sum_{j=1}^{k'} c_j' n_1' \cdots n_j'$$

where $\qquad c_j' = c_j \; (j = 1, \cdots, k' - 1)$

and

$$c_{k'}' = c_{k'} + \sum_{l=1}^{k-k'} c_{k'+l} n_{k'+1} \cdots n_{k'+l} \tag{27}$$

$$h_2(n_1', \cdots, n_{k'}') = \sum_{j=1}^{k'} \frac{\sigma_j^2}{n_1' \cdots n_j'} + \frac{1}{n_1' \cdots n_{k'}'} \sum_{l=1}^{k-k'} \frac{\sigma_{k'+l}^2}{n_{k'+1} \cdots n_{k'+l}} \tag{28}$$

$$= \sum_{j=1}^{k'} \frac{\sigma_j'^2}{n_1' \cdots n_j'}$$

where $\qquad \sigma_j' = \sigma_j \; (j = 1, \cdots, k' - 1)$

and

$$\sigma_{k'}'^2 = \sigma_{k'}^2 + \sum_{l=1}^{k-k'} \frac{\sigma_{k'+l}^2}{n_{k'+1} \cdots n_{k'+l}} \tag{29}$$

Thus the functions h_1 and h_2 of the variables $n_1', \cdots, n_{k'}'$, given by (26) and (28), represent the same types of function as the functions g_1 and g_2 of the variables $n_1, \cdots, n_k$ given by (22) and (23). Therefore the minimum solutions of $h_1(n_1', \cdots, n_{k'}')$ and $h_2(n_1', \cdots, n_{k'}')$ under the side conditions $h_2(n_1', \cdots, n_{k'}') = v$ and $h_1(n_1', \cdots, n_{k'}') = c$ respectively, may be obtained from equations (24) and (25) by replacing k by k', σ by σ', and c by c', and then substituting back σ_j' and c_j' $(j = 1, \cdots, k')$ from equations (27) and (29).

For $k = 3$, $k' = 2$ we obtain from (27) and (29)

$$c_1' = c_1 \qquad c_2' = c_2 + c_3 n_3$$

$$\sigma_1' = \sigma_1 \qquad \sigma_2'^2 = \sigma_2^2 + \frac{\sigma_3^2}{n_3}$$

The substitution of these values into (24) and (25) after replacement of k, c, σ by k', c', σ' gives the formulas (9) and (10) used above.

Note that the results of b. ii. may also be obtained from a. and then b. i. be considered as the special case $k' = k$.

iii. *Stratified Sampling from Finite Populations*

We will indicate briefly the applicability of the above used generalized functions to stratified sampling involving two levels.

Let there be k strata in the population with M_i elements x_{ij} in the i-th stratum ($i = 1, \cdots, k; j = 1, \cdots, M_i$). Assume that the N_i sample elements x_{ij} ($i = 1, \cdots, k; j = 1, \cdots, N_i$) are independently drawn at random from the k finite strata. Then the sample mean

$$\bar{x} = \frac{1}{M} \sum_{i=1}^{k} M_i \frac{\sum_{j=1}^{N_i} x_{ij}}{N_i}$$

has the variance

$$\sigma_{\bar{x}}^2 = \frac{1}{M^2} \sum_{i=1}^{k} M_i^2 \frac{\sigma_i^2}{N_i} \frac{M_i - N_i}{M_i - 1}$$

where $M = \sum_{i=1}^{k} M_i$ and σ_i^2 denotes the variance between elements in the i-th stratum. Thus we have

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{k} \frac{a_{2i}}{N_i} + a_2$$

where $\quad a_{2i} = \dfrac{M_i^3 \sigma_i^2}{M^2(M_i - 1)} \quad$ and $\quad a_2 = -\dfrac{1}{M^2} \displaystyle\sum_{i=1}^{k} \dfrac{M_i^2 \sigma_i^2}{M_i - 1}$

Let c_i be the cost per element in the i-th stratum and $c = \sum_{i=1}^{k} c_i N_i$ the total cost, then c may be written $c = \sum_{i=1}^{k} a_{1i} N_i + a_1$ where $a_{1i} = c_i$ and $a_1 = 0$. Thus c and $\sigma_{\bar{x}}^2$ correspond to the functions $F_1(N_1, \cdots, N_k)$ and $F_2(N_1, \cdots, N_k)$ respectively in (11) and (12). Therefore equations (14) and (16) give the desired minimum solutions where b_1 and b_2 determine the side conditions corresponding to (13) and (15). In case the populations in the strata are large ($M_i \sim M_i - 1$), we obtain the well known optimum allocation formulas:

$$N_{1i} = \frac{\sum_{i=1}^{k} (M_i \sigma_i \sqrt{c_i})}{M^2 b_2 + \sum_{i=1}^{k} (M_i \sigma_i^2)} \frac{M_i \sigma_i}{\sqrt{c_i}}$$

$$N_{2i} = \frac{b_1}{\sum_{i=1}^{k} M_i \sigma_i \sqrt{c_i}} \frac{M_i \sigma_i}{\sqrt{c_i}}$$

LITERATURE CITED

[1] Anderson, R. L. Use of Variance Components in the Analysis of Hog Prices in Two Markets, *J. Am. Stat. Ass.*, 42: 612–634, 1947.

[2] Cassil, C. C., Wadley, F. M., and Dean, F. P. Sampling Studies on Orchard Spray Residues in the Pacific Northwest, *J. of Econ. Entom.*, 36: 227–231, 1943.

[3] Crump, S. Lee. The Estimation of Variance Components in Analysis of Variance, *Biometrics Bulletin*, 2: 7–11, 1946.

[4] Daniels, H. E. The Estimation of Components of Variance, *Supplement to the Journal of the Royal Statistical Society*, 6: 186–197, 1939.

[5] Eisenhart, Churchill. The Assumptions Underlying the Analysis of Variance, *Biometrics Bulletin*, 3: 1–21, 1947.

[6] Fisher, R. A. *The Design of Experiments*, 3rd Edition. Oliver and Boyd, Ltd., Edinburgh and London, 1942.

[7] Ganguli, M., A Note on Nested Sampling, *Sankhya*, 5: 449–452, 1941.

[8] Hansen, M. H., Hurwitz, W. N., and Gurney, M. Problems and Methods in a Sample Survey of Business, *J. Am. Stat. Ass.*, 41: 173–189, 1946.

[9] Mahalanobis, P. C. On Large Scale Sample Surveys, *Philos, Transactions of the Royal Society, Series B, Biolog. Sciences*, 23: 329–451, 1944.

[10] Snedecor, G. W. *Statistical Methods Applied to Experiments in Agriculture and Biology*, 4th Edition. The Collegiate Press, Inc., Ames, Iowa, 1946.

[11] Statistical Research Group, Columbia University. *Selected Techniques of Statistical Analysis: For Scientific and Industrial Research and Production and Management Engineering.* McGraw-Hill Book Company, Inc., New York, 1947.

[12] Tippett, L. H. C. *The Methods in Statistics*, 3rd Edition. Williams and Norgate, Ltd., London, 1940.

FITTING A STRAIGHT LINE WHEN BOTH VARIABLES ARE SUBJECT TO ERROR

M. S. Bartlett
University of Manchester, England

INTRODUCTION

A simple method of fitting a straight line when both variables are subject to error was examined by Wald (1) in 1940. The purpose of the present note is to present and illustrate a modification of Wald's method having the advantage in general of greater accuracy. Before any detailed exposition it will be as well to recall two important points:

(i) a distinction must be made between the linear regression equation of a variable y on a second variable x, and a linear functional relation between two variables Y and X masked by errors. The former equation is still available for prediction even if the variable x is subject to error, but is not necessarily appropriate for a functional relation when one exists.

(ii) it is possible to set up maximum likelihood equations for the second problem, but they do not lead to a unique solution without further assumptions, such as an assumption about the relative magnitude of the errors in x and y.

These points have been emphasized by many previous writers, for example, by Wald (1) or more recently by Lindley (2). In view of (ii) it is useful to consider, in the common case when the observations have equal weight, the following elementary method:

(a) For the location of the fitted straight line use as one point the mean coordinates $\bar{x}$, $\bar{y}$, just as in the least-squares method.

(b) For the slope, first divide the n plotted points into three groups, the equal numbers k in the two extreme groups being chosen to be as near $\frac{1}{3}n$ as possible (the three groups are non-overlapping when considered, say, in the x direction). The join of the mean coordinates $\bar{x}_1$, $\bar{y}_1$ and $\bar{x}_3$, $\bar{y}_3$ for the two extreme groups is used to determine the slope.

The only difference from Wald's original method is the use of three groups instead of two, for reasons which will be apparent from the results of the next section.[1] It will also be shown that Wald's confidence interval method of assessing the accuracy (under suitable conditions) may be adapted to the present method.

EFFICIENCY IN A SPECIAL CASE

To get some idea of the efficiency of the method its accuracy is determined in a special case where least-squares is appropriate. It is assumed that observations y are available for $n = 2l + 1$ values $x \equiv X$ not subject to error and spaced at equidistant unit intervals. The least-squares estimate is known to provide the linear combination of the y's providing an unbiased estimate of the true slope β in the functional relation

$$Y = \alpha + \beta X \tag{1}$$

with minimum variance when the differences $y - Y$ are uncorrelated and of constant variance σ^2. The least-squares estimate

$$b = \sum y(x - \bar{x}) / \sum (x - \bar{x})^2$$

has error variance $\sigma^2 / \sum (x - \bar{x})^2$, where $\sum (x - \bar{x})^2 = (\frac{1}{3})l(l + 1)(2l + 1)$ in the situation assumed in this section.

For comparison the error variance of the estimate

$$b' = \frac{\bar{y}_3 - \bar{y}_1}{\bar{x}_3 - \bar{x}_1} \tag{2}$$

of the last section is easily evaluated for any value of k. It is given by

$$\frac{2\sigma^2}{k(\bar{x}_3 - \bar{x}_1)^2} = \frac{2\sigma^2}{k(2l - k + 1)^2} \cdot$$

The relative efficiency of b' is thus

$$E = \frac{3k(2l - k + 1)^2}{2l(l + 1)(2l + 1)} \cdot \tag{3}$$

This is a maximum when

$$(2l - k + 1)(2l + 1 - 3k) = 0$$

with relevant root $k = (\frac{1}{3})(2l + 1) = \frac{1}{3}n.$

[1] I am indebted to Professor Gerhard Tintner for drawing my attention to a previous discussion of this problem, with a similar conclusion, by Nair and Shrivastava (4)(see also Nair and Banerjee (5)). It might be noted that these authors propose using the two extreme groups out of three for location as well as slope, but recommendation (a) above is theoretically preferable. In the first of these two papers the extension of the method to fitting higher-order curves is also considered, though the optimum efficiency is not so high in such cases.

We then have

$$E = \frac{8(l + \frac{1}{2})^2}{9l(l + 1)} \geq \frac{8}{9}, \tag{4}$$

which may be compared with $E = (\frac{3}{4})(l + \frac{1}{2})^2/[l(l + 1)] \geq 3/4$ when $k = \frac{1}{2}n$. The higher efficiency of $k = \frac{1}{3}n$ compared with $k = \frac{1}{2}n$ suggests the adoption of $k = \frac{1}{3}n$ in preference to $k = \frac{1}{2}n$ in general. Indeed its high efficiency in the case examined above indicates the occasional value of the simple method proposed even in cases where the least-squares method is available.

ASSESSMENT OF ACCURACY IN THE GENERAL CASE

In the general problem it is assumed that both y and x are subject to error. To use Wald's confidence interval method it is assumed further that the n errors $\eta \equiv y - Y$ are independently and normally distributed with constant variance σ_η^2, similarly the n errors $\epsilon \equiv x - X$ are independent and normal with variance σ_ϵ^2; the x and y errors are moreover mutually independent, so that the variance of $\eta - \beta\epsilon$ is $\sigma_\eta^2 + \beta^2\sigma_\epsilon^2$.

Consider now possible 'estimates' of this last variance when β is known. If we write for the total sums of squares and products of x and y *within* the three groups

$$S_{xx} \equiv \sum\nolimits_1 (x - \bar{x}_1)^2 + \sum\nolimits_2 (x - \bar{x}_2)^2 + \sum\nolimits_3 (x - \bar{x}_3)^2$$

$$S_{xy} \equiv \sum\nolimits_1 (x - \bar{x}_1)(y - \bar{y}_1) + \sum\nolimits_2 (x - \bar{x}_2)(y - \bar{y}_2) + \sum\nolimits_3 (x - \bar{x}_3)(y - \bar{y}_3)$$

$$S_{yy} \equiv \sum\nolimits_1 (y - \bar{y}_1)^2 + \sum\nolimits_2 (y - \bar{y}_2)^2 + \sum\nolimits_3 (y - \bar{y}_3)^2,$$

where $\sum_i$ denotes summation over the observations in the i-th group, then $(S_{yy} - 2\beta S_{xy} + \beta^2 S_{xx})/(n - 3)$ is an estimate of the variance $\sigma_\eta^2 + \beta^2\sigma_\epsilon^2$ with $n - 3$ degrees of freedom. The remaining 3 degrees of freedom are contained in the three group means. One is represented by the general mean, one by the difference between the means of the first and third groups to be used in the estimate of β; the third is represented by the difference between the mean of the second group and the general mean of the first and third groups.

For data with few observations it is advisable to make use of the last degree of freedom in the variance estimate, as in the numerical example considered later. Alternatively, if it is not so used, it remains available for testing the linearity of the true X, Y relation. In the former case, the appropriate square to be added to the numerator of the previous estimate is

$$\{(\bar{y}_1 + \bar{y}_3 - 2\bar{y}_2)^2 - 2\beta(\bar{y}_1 + \bar{y}_3 - 2\bar{y}_2)(\bar{x}_1 + \bar{x}_3 - 2\bar{x}_2) + \beta^2(\bar{x}_1 + \bar{x}_3 - 2\bar{x}_2)^2\}\left\{\frac{2}{k} + \frac{4}{n-k}\right\}$$

and the estimate $s^2(\beta)$ obtained with $n - 2$ now as the divisor will have $n - 2$ degrees of freedom.

Since

$$(\bar{x}_3 - \bar{x}_1)(b' - \beta) = (\bar{\eta}_3 - \beta\bar{\epsilon}_3) - (\bar{\eta}_1 - \beta\bar{\epsilon}_1),$$

when b' is given by (2), the left-hand quantity under the assumptions made in this section is normal with variance $(\sigma_\eta^2 + \beta^2\sigma_\epsilon^2)(2/k)$. This is subject to one qualification, that the errors in the x variable do not influence the allocation of the observations to the three groups. Such an effect may be neglected in many problems, particularly when the errors are small compared with the spacing of the observations at the points of division between the three groups; it will not be considered further here. A more detailed consideration of this point has been given by Wald (1).

Under the same assumptions we have

$$t = \frac{(\bar{x}_3 - \bar{x}_1)(b' - \beta)\sqrt{\tfrac{1}{2}k}}{s(\beta)} .$$

Although the denominator depends on β, this t-variate enables a confidence interval to be obtained for β. Thus for a value t corresponding to any chosen probability value we have the interval determined by the quadratic equation for β,

$$(5) \qquad (\bar{x}_3 - \bar{x}_1)^2(b' - \beta)^2\tfrac{1}{2}k = t^2(s_y^2 - 2\beta s_{xy} + \beta^2 s_x^2),$$

where $s^2(\beta) \equiv s_y^2 - 2\beta s_{xy} + \beta^2 s_x^2$.

If required, a similar method may be used to provide a joint confidence region for α and β. If $a \equiv \bar{y} - \beta\bar{x}$, then a is independent of the numerator of t and of $s(\beta)$, and hence

$$F = \frac{\tfrac{1}{2}\{n(a - \alpha)^2 + \tfrac{1}{2}k(\bar{x}_3 - \bar{x}_1)^2(b' - \beta)^2\}}{s^2(\beta)}$$

is a variance ratio with degrees of freedom 2, $n - 2$. For any chosen probability value the corresponding critical value of F will determine an ellipse as the boundary of the confidence region for α and β. This may be compared with the corresponding region for the least-squares method if it is known that $\sigma_\epsilon^2 = 0$; this region is similarly obtained from the variance ratio

$$F = \frac{\frac{1}{2}\{n(a-\alpha)^2 + (b-\beta)^2 \sum (x-\overline{x})^2\}}{s^2},$$

where s^2 is the usual variance estimate of $y - Y$ obtained from the residuals of y with $n - 2$ degrees of freedom.

If, as suggested earlier in this section, it is desired to examine the linearity of the functional relation, the variance estimate $s^2_{n-3}(\beta)$ of $\sigma_\eta^2 + \beta^2\sigma_\epsilon^2$ with $n - 3$ degrees of freedom must be used. The further quantity

$$t = \frac{\{(\overline{y}_1 + \overline{y}_3 - 2\overline{y}_2) - \beta(\overline{x}_1 + \overline{x}_3 - 2\overline{x}_2)\}\left\{\frac{2}{k} + \frac{4}{n-k}\right\}^{-\frac{1}{2}}}{s_{n-3}(\beta)}$$

is then (if the linear relation is valid) also a t-variate with $n - 3$ degrees of freedom. It will be noticed that it involves the unknown slope β. When this is replaced by the estimate b', the resulting statistic is no longer exactly a t-variate, but might be treated approximately as such, especially when $\overline{x}_1 + \overline{x}_3 - 2\overline{x}_2$ is small compared with $\overline{x}_3 - \overline{x}_1$.

NUMERICAL EXAMPLE

As a numerical example consider fitting a straight line to the data on penicillin 'assay' given by Davies (3, S 6.12). Six different concentrations of pure penicillin were set up on a plate on which an agar medium containing *B. subtilis* had been spread, and the mean circle diameters of the zones of inhibition of growth of the organisms were measured (for further details of the technique see S 5.41 of (3)). The concentration had negligible error, so that the standard least-squares method was available, the relation between circle diameter and log. concentration being linear. With circle diameter y in mms. and 1 penicillin unit per ml. as $x = 1$, and a two-fold increase in concentration as the unit for the x scale, the regression equation of y on x was

$$Y = 20.403 + 1.782(x - 3.5) = 14.166 + 1.782\,x \tag{6}$$

with a 95% confidence interval for the slope, based on the usual t-statistic, of (1.732, 1.832).

It is stressed that the data are considered again here purely in order to illustrate the present method. The six observations are divided into three groups:

y	15.87	17.78,	19.52	21.35,	23.13	24.77	(Total 122.42)
x	1	2 ,	3	4 ,	5	6	(Total 21)

$$b' = \frac{(24.77 + 23.13) - (17.78 + 15.87)}{(6 + 5) - (2 + 1)} = 1.781.$$

Hence the estimated relation is

$$(7) \qquad Y = 20.403 + 1.781(X - 3.5) = 14.170 + 1.781X.$$

The sum of squares within each group has only one degree of freedom in this example, and may conveniently be calculated from the difference of the two observations per group. The other degree of freedom to be added is that for the contrast of the mean for the second group with the mean for the other two groups. This gives zero contribution for x, and for y

$$24.77 + 23.13 + 15.87 + 17.78 - 2(19.52 + 21.35) = -0.19$$

with appropriate divisor. Hence

$$4s_y^2 = \frac{(1.91)^2 + (1.83)^2 + (1.64)^2}{2} + \frac{(-0.19)^2}{12} = 4.8463$$

$$4s_{xy} = \frac{1 \times 1.91 + 1 \times 1.83 + 1 \times 1.64}{2} + \frac{(0 \times -0.19)}{12} = 2.69$$

$$4s_x^2 = \frac{1^2 + 1^2 + 1^2}{2} + \frac{0^2}{12} = 1.5.$$

Equation (5), with $t = 2.78$ for 4 degrees of freedom ($P = 0.05$), gives

$$16(1.781 - \beta)^2 = (2.78)^2(4.8463 - 2\beta \times 2.69 + 1.5\beta^2)/4$$

or
$$13.1018\beta^2 - 2\beta(23.2987) + 41.3879 = 0$$

or
$$\beta = 1.778 \pm 0.058.$$

Thus the 95% confidence interval for β by this method is (1.720, 1.836), an interval naturally slightly wider than the interval obtained by the least-squares method, since the assumption of no error in x has been dropped.

REFERENCES

(1) Wald, A. The Fitting of Straight Lines if Both Variables are Subject to Error. *Ann. Math. Stat.* 11, 284, 1940.

(2) Lindley, D. V. Regression Lines and the Linear Functional Relationship. *J. Roy. Stat. Soc.* (*Suppl.*) 9, 218, 1947.

(3) Davies, O. L. (Editor). *Statistical Methods in Research and Production.* Oliver and Boyd, 1947.

(4) Nair, K. R. and Shrivastava, M. P. On a Simple Method of Curve Fitting. *Sankhyā* 6, 121, 1942.

(5) Nair, K. R. and Banerjee, K. S. A Note on Fitting of Straight Lines if Both Variables are Subject to Error. *Sankhyā* 6, 331, 1942.

RELATIONSHIP OF CATCH TO CHANGES IN POPULATION SIZE OF NEW ENGLAND HADDOCK

By Howard A. Schuck

Aquatic Biologist

Fish and Wildlife Service
United States Department of the Interior

INTRODUCTION

The United States catch of haddock has fluctuated considerably throughout the years and these fluctuations have generally been of a declining nature. In 1929, the catch was about 260 million pounds, and in recent years it has averaged only about 150 million pounds. Fluctuations in the catch have been due in large part to variations in actual abundance, or the size of the stock of commercial sizes of haddock on the banks in different years. We are, therefore, interested in obtaining an accurate measure of the size and the composition of the stock, to measure its changes throughout the years, and to determine what factors have been most responsible for such changes. Changes in the stock from year to year are the result of varying rates of removals and additions. Therefore, besides determining the size of the stock in different years, it is necessary to measure the yearly removals from the stock by catch and natural mortality, and the yearly additions by recruitment and growth.

If these variables could be measured accurately, we should be in a position to evaluate their relative importance in determining the size of the stock and to determine whether the size of the spawning stock and of other stocks affects recruitment. With such information and other general life history facts, it should be possible to determine at what level the stock should be maintained, to determine what mode and intensity of fishing will result in the maximum sustained production of haddock, and to make periodic predictions as to future production of the fishery for the industry.

A basic equation is:

$$S + (G + R + M) - (C + N + M_1) = S_1$$

where:

S = size of population at the beginning of the year.
S_1 = size of population at the end of the year.
G = additions to the population during the year by growth.
R = additions to the population during the year by recruitment of young.
M = additions due to immigrations.
C = deductions from the population during the year by the fishery.
N = deductions from the population during the year due to natural mortality.
M_1 = deductions due to emigrations.

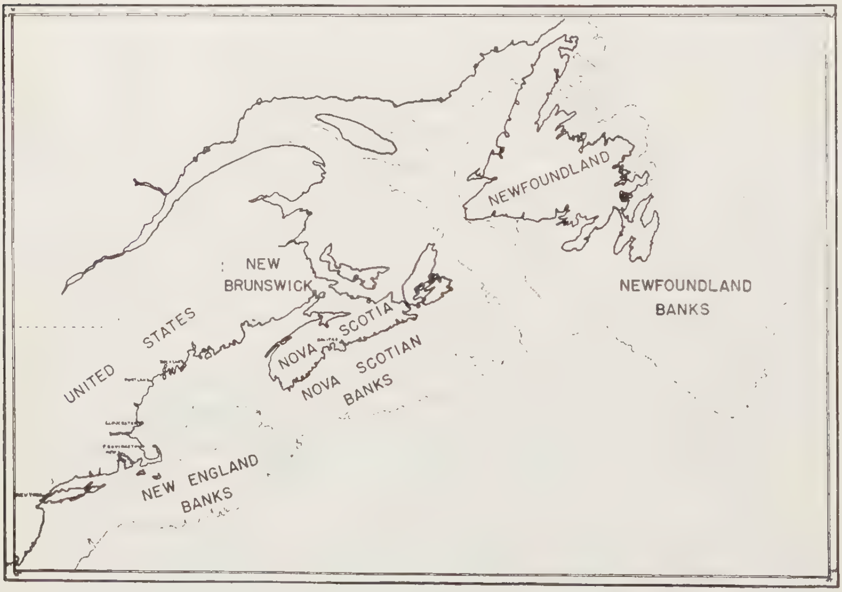


FIGURE 1.
LOCATION OF FISHING BANKS OFF NEW ENGLAND, NOVA SCOTIA, AND NEWFOUNDLAND.

It is believed that the population of haddock inhabiting the New England Banks (Georges) (Fig. 1) is largely independent of the populations on the Nova Scotian and Newfoundland Banks. Assuming this to be true, and if we consider the population on Georges Bank only, there will be no important changes in the stock from year to year due to

TABLE 1

RELATIVE SIZE OF THE GEORGES BANK HADDOCK POPULATIONS IN TERMS OF THE AVERAGE NUMBERS OF FISH PER DAY TAKEN BY A STANDARD GROUP OF OTTER TRAWLERS

Year	Numbers per day
1931	3,032
1932	4,324
1933	3,630
1934	4,049
1935	4,927
1936	5,590
1937	4,404
1938	4,833
1939	5,502
1940	4,979
1941	6,960
1942	7,941
1943	7,319
1944	5,737
1945	5,347
1946	4,956
1947	4,954
Average	5,205

immigrations or emigrations; and M amd M_1 can be left out of the equation.

Also, if we consider the population as numbers, rather than pounds of fish, G or "growth", can be left out too. Furthermore, if we define the population S as being the number of fish of certain year classes at the beginning of a year and S_1 as the number of fish of the same year classes at the end of that year, then there can be no recruitment; and R can also be ignored. Thus, the equation for certain purposes can be reduced to:

$$S - (C + N) = S_1$$

Available for use in this equation are biological and statistical data for the Georges Bank population going back to 1931. These data were assembled by the Haddock Investigation of the United States Fish and Wildlife Service and its predecessor agency, the United States Bureau of Fisheries.

The remainder of this paper will be devoted to: (1) developing an index representing the size of the population in terms of numbers of

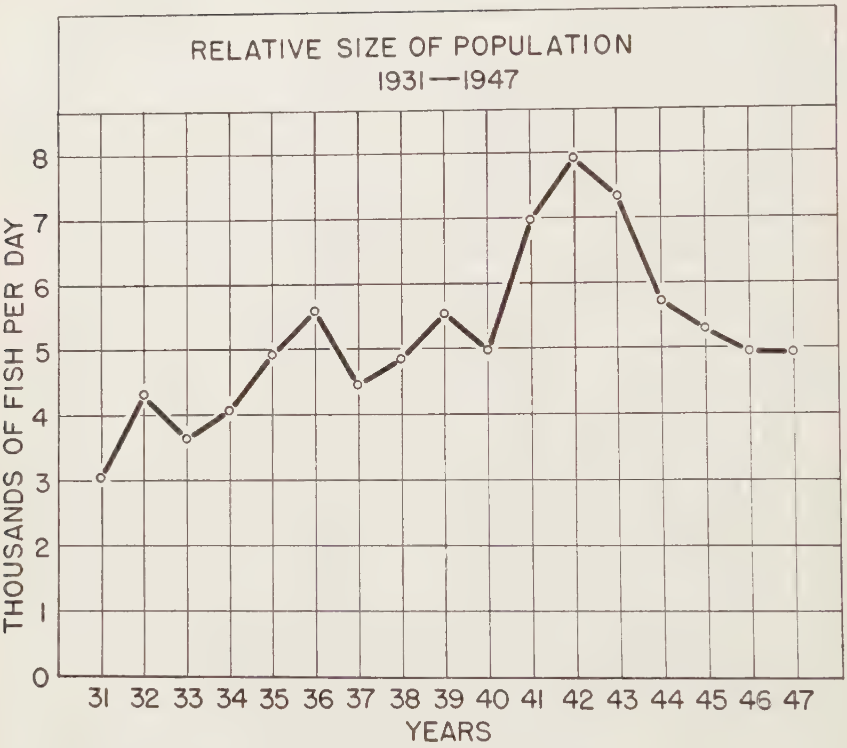


FIGURE 2.
RELATIVE SIZE OF THE POPULATION, IN TERMS OF THOUSANDS OF FISH PER DAY BY YEARS.

haddock of definite ages and year classes, at the beginning and end of yearly periods (S and S_1); (2) measuring the fishery removals (C) of haddock of each age during each of the 17 years, 1931–47; and (3) determining how important the yearly fishery removals are in decreasing the stock from the beginning to the end of yearly periods.

SIZE OF THE STOCK OR "S" AND "S_1"

Total catch represents fishing removals and in itself is a vital piece of information. It does not, however, represent abundance, or the relative size of the population on the Bank, inasmuch as the amount of fishing effort utilized to make the catch varies among years.

The index representing the relative size of the population that was used in the Haddock Investigation is the average yearly catch per day[1] of a standard group of large otter trawlers which fished out of Boston

[1] Details of this abundance analysis were developed by W. C. Herrington and G. A. Rounsefell.

during this 17-year period. The relative size of the population[2] was first expressed in terms of the average number of pounds per day taken by these trawlers in each year. By the use of yearly average weight data, the statistics on relative population size were converted from pounds to numbers of fish (Table 1 Fig. 2).

In each year and season a *sample* of the haddock that were landed had been obtained, and from those fish obtained, scales had been collected. Then, for each year and season, the ages of these sample fish were determined. This determination was made by studying the projected impression of these scales. Figure 3 shows a photograph of such a microprojection of a scale from a Georges Bank haddock.

The fish were aged as having completed their first, second, third, fourth, fifth, sixth, seventh, eighth, and ninth year of life, and were correspondingly classified as fish of 1 to 9 years of age. The category of 9-year-olds includes 9-year-old and older fish. (The number of haddock of ages greater than 9 years was very small, amounting in the aggregate to less than one-half of one percent of all haddock in the catch.)

By using the percentage age composition that had been computed for haddock of each length and for each year and season, the total numbers of fish caught per day were reduced to numbers per day of each age (Table 2 and Fig. 4). The average abundance for all 17 years (Fig. 5) amounted to 116 one-year-olds, 1,472 two-year-olds, 1,571 three-year-olds, 920 four-year-olds, 557 five-year-olds, 324 six-year-olds, 149 seven-year-olds, 61 eight-year-olds and 34 nine-year-old and older fish. It can be concluded that the relative abundance of fish in the catch diminishes quite regularly for those fish three years old and older. The fact that the one- and two-year-old fish are less abundant indicates that these age groups are not fully available to the fishery.

In order to measure the diminution in the stock over the period of a year, it was desired to compare the relative population size of the fully-available age groups of fish at the beginning of each year with the size of the corresponding stock at the end of the year. Table 2 gives the average population size for the "haddock" year.[3] In order to obtain a

[2] Where "population size" is mentioned in the remainder of this paper it refers to this index of relative population size. Although it has not yet been possible to determine the exact relationship between the actual number or pounds of fish in the stock and our calculated index of relative population size, the index appears to suffice for the purpose used here.

[3] The "haddock" year consists of seasons *A*, *B*, *C*, and *D* as follows:

A—February, March, and April (spawning season)
B—May, June, and July
C—August, September, and October
D—November, December, and January

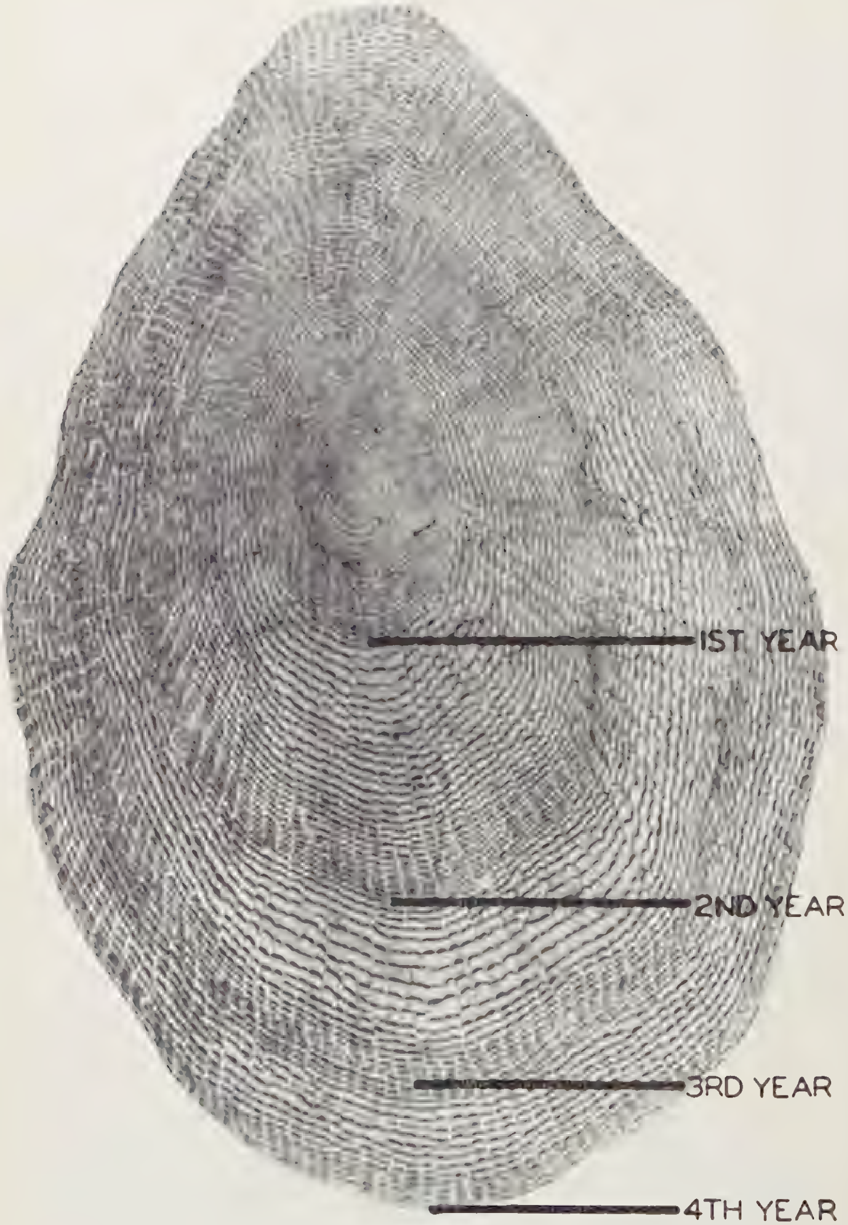


FIGURE 3.

PHOTOGRAPH OF A SCALE FROM A HADDOCK THAT HAD JUST COMPLETED ITS FOURTH YEAR WHEN CAUGHT APRIL 1939 ON GEORGES BANK. THE MARKS INDICATE THE COMPLETION OF EACH YEAR OF GROWTH. THE LENGTH OF THIS FISH WAS 22 3/4 INCHES.

TABLE 2

RELATIVE POPULATION OF EACH AGE OF GEORGES BANK HADDOCK IN TERMS OF NUMBERS CAUGHT PER DAY

Year	Age in years									
	All Ages	1	2	3	4	5	6	7	8	9 and older
1931	3,032	147	691	158	439	699	466	256	132	44
1932	4,324	11	210	2,829	275	413	323	146	74	43
1933	3,630	44	986	720	1,145	249	193	145	67	81
1934	4,049	141	966	1,108	690	678	241	125	60	40
1935	4,927	202	1,704	1,306	574	509	428	97	74	33
1936	5,590	157	1,752	1,834	920	402	236	222	41	26
1937	4,404	150	1,233	1,327	698	535	251	119	65	26
1938	4,833	165	2,590	988	489	234	199	114	31	23
1939	5,502	95	1,775	2,416	640	271	123	108	42	32
1940	4,979	524	1,116	1,689	1,018	309	184	93	28	18
1941	6,960	144	3,298	1,275	1,046	752	233	123	40	49
1942	7,941	94	3,036	2,567	1,037	624	362	158	36	27
1943	7,319	11	1,026	3,470	1,551	530	492	149	61	29
1944	5,737	14	135	1,412	2,609	948	416	95	91	17
1945	5,347	25	1,663	420	1,244	1,218	485	194	61	37
1946	4,956	24	856	1,992	400	854	562	217	49	2
1947	4,954	18	1,996	1,189	863	250	314	180	89	55
Total	88,484	1,966	25,033	26,700	15,638	9,475	5,508	2,541	1,041	582
Avg.	5,205	116	1,472	1,571	920	557	324	149	61	34

value of the population size at the beginning of the year while eliminating the effect of the seasonal cycle in availability it was necessary to recompute these data.

All data originally had been computed on a seasonal basis: for example, Table 3 shows the seasonal population size data from which Table 2 was derived. In order to obtain values for the population size that more closely represent values at the beginning of each year,[4] the abundance values for seasons *C*, *D*, *A*, and *B* in Table 3 were averaged. In this recombination it was necessary to consider that 3-year-old fish in seasons *C* and *D* become 4-year-old fish in seasons *A* and *B* of the following year and that other ages progress accordingly.

For example, to obtain the relative size of the population of 4-year-old fish at the beginning of 1935 the following figures[5] were used:

[4] It is recognized that by summarizing values for 4 seasons and dividing by 4, the average obtained does not under some conditions represent the average of the midpoint and thus the exact beginning of the year. For the purpose of this analysis, however, such a calculation represents the beginning of the year accurately enough.

[5] An exception to this rule was made in computing the relative population of 9-year-old haddock at the beginning of the year. Since this group includes all older haddock, 8-year-old and 9-year-old haddock from seasons *C* and *D* were added to 9-year-old haddock from seasons *A* and *B* and the total of these 6 figures, instead of the usual 4, was divided by 4 to give the average.

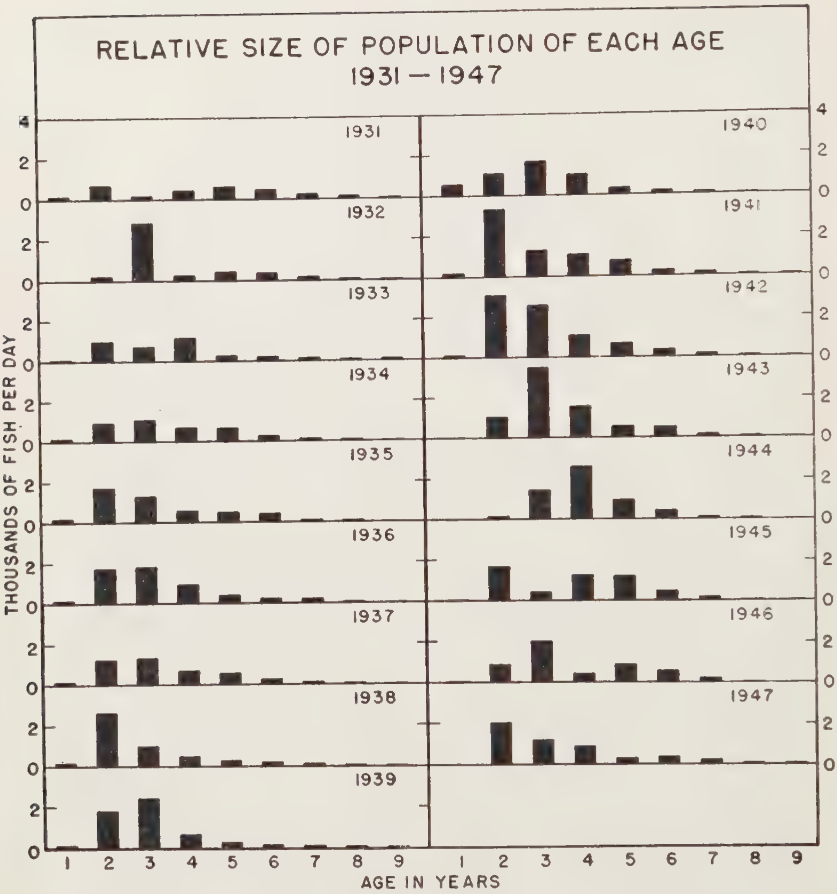


FIGURE 4.

RELATIVE SIZE OF HADDOCK POPULATION OF AGES 1–9, FOR EACH OF THE 17 YEARS.

	Number of fish per day
3-year-old haddock, season *C*, 1934	1,425
3-year-old haddock, season *D*, 1934	431
4-year-old haddock, season *A*, 1935	583
4-year-old haddock, season *B*, 1935	889
Total	3,328
Average	832

Using this system the relative sizes of the population of 4- to 9-year-old fish at the beginning of each year were computed and are shown in

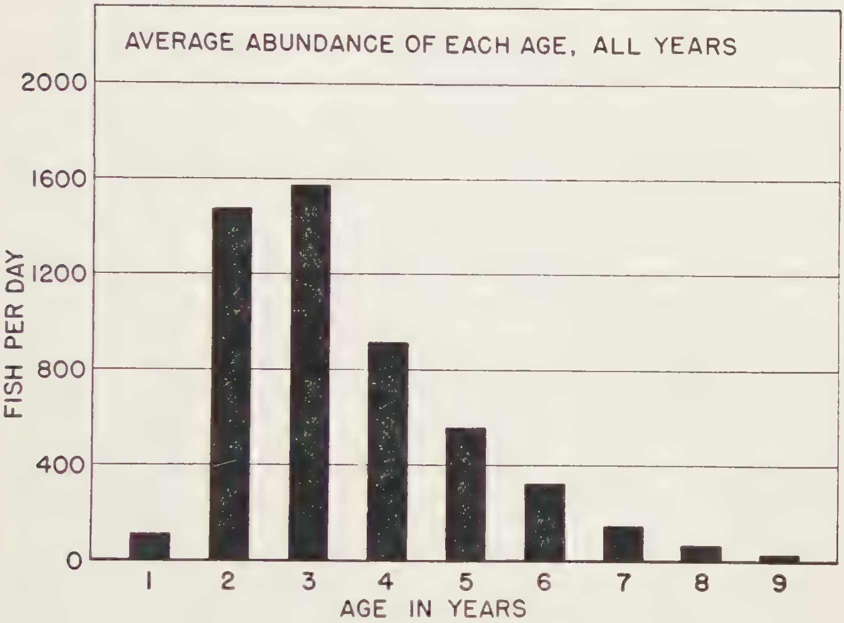


FIGURE 5.
RELATIVE POPULATION SIZE OF HADDOCK OF AGES 1–9. AVERAGE OF ALL 17 YEARS.

Table 4. Since computation of the catch-per-day of 3-year-old fish at the beginning of the year involved use of figures for the less available 2-year-old fish in seasons C and D, it was decided to omit the 3-year group.

The next step was to decide whether to consider the age groups separately or in the aggregate. Examination of the data in Table 3 indicated that the decrease in catch-per-day for individual year classes from year to year was rather variable, hence, it was desirable to combine age groups. Thus the total of all fish of ages 4 to 9 years for the beginning of each year are shown in the right-hand column of Table 4.

It was next necessary to compute the size of the population at the end, in addition to at the beginning, of each year. The average population at the beginning of the year or seasons $(C + D + A + B)/4$, approximates the value of the midpoint between D and A. Therefore, values for the number of fish at the beginning of the year are the same as values for the number at the *end* of the preceding year.

For example, from Table 4, if there are 1,793 five-year-old fish per day at the beginning of 1945, then there are 914 (the number of 6-year-olds at the beginning of 1946) 5-year-olds at the end of 1945.

TABLE 3

RELATIVE SIZE OF POPULATION OF EACH AGE OF GEORGES BANK HADDOCK BY SEASONS IN NUMBERS CAUGHT PER DAY

Year	Season	Age in years									
		No. all years	1	2	3	4	5	6	7	8	9 and older
1931	*A*	3,268	...	30	193	560	1,088	781	372	155	89
	B	3,182	...	81	70	770	1,139	630	336	119	37
	C	2,562	...	897	186	245	348	353	269	224	40
	D	3,114	587	1,755	183	179	222	99	45	32	12
1932	*A*	4,281	...	11	2,418	489	707	387	143	80	46
	B	4,937	...	38	3,333	359	410	351	275	95	76
	C	5,848	3	430	4,253	149	343	458	96	91	25
	D	2,231	41	361	1,310	103	191	96	70	32	27
1933	*A*	3,697	...	112	254	1,728	411	423	324	183	262
	B	4,349	...	1,318	494	1,814	277	185	161	56	44
	C	4,487	39	1,724	1,461	875	207	93	63	17	8
	D	1,988	138	789	671	164	100	72	31	12	11
1934	*A*	3,729	...	4	1,360	471	874	574	198	150	98
	B	4,299	...	290	1,217	1,368	866	209	252	41	56
	C	4,619	...	1,929	1,425	544	502	148	33	37	1
	D	3,549	565	1,640	431	377	472	33	17	10	4
1935	*A*	3,215	...	23	884	583	748	607	185	86	99
	B	5,536	...	1,117	1,719	889	740	884	52	131	4
	C	5,495	16	2,443	1,848	590	367	152	67	9	3
	D	5,462	791	3,235	773	234	179	71	84	71	24
1936	*A*	5,827	...	267	2,078	1,688	903	397	356	47	91
	B	7,217	4	1,760	3,118	997	517	341	377	103	...
	C	6,171	93	3,618	1,537	692	41	69	110	2	9
	D	3,143	532	1,361	603	303	145	138	45	11	5
1937	*A*	5,224	1	423	1,810	1,167	988	500	176	137	22
	B	4,969	...	1,068	1,858	949	552	282	161	54	45
	C	5,175	361	2,757	1,265	439	243	24	51	21	14
	D	2,247	237	685	375	237	356	196	90	46	25
1938	*A*	3,078	...	363	1,015	694	341	390	159	56	60
	B	4,736	...	2,133	1,233	614	288	270	178	13	7
	C	7,425	...	5,500	1,241	384	162	54	52	26	6
	D	4,092	660	2,363	463	264	144	81	68	28	21
1939	*A*	4,463	...	287	2,260	873	515	191	179	66	92
	B	6,629	...	1,604	3,371	958	251	223	143	66	13
	C	6,848	30	3,057	3,157	424	148	16	13	3	...
	D	4,066	349	2,150	877	307	174	56	96	34	23

Also, the aggregate population of 4-year-old and older fish at the beginning of a year would produce survivors at the end of the year which would amount to the number of 5-year and older fish at the beginning of the next year. For example, if the total population of 4-year-old and older fish at the beginning of 1944 was the total of 3,188 four-year-

TABLE 3—*Continued*

Year	Season	Ages in years									
		No. all years	1	2	3	4	5	6	7	8	9 and older
1940	*A*	2,805	...	127	1,057	1,046	337	141	52	24	21
	B	6,245	...	1,200	2,449	1,521	509	234	268	42	22
	C	5,638	221	2,175	1,832	922	161	289	10	23	5
	D	5,228	1,876	961	1,415	582	230	73	43	24	24
1941	*A*	5,855	1	1,463	1,735	1,289	1,003	181	135	22	26
	B	7,692	...	2,165	1,732	1,639	1,222	497	209	96	132
	C	9,082	128	6,498	1,075	757	381	123	81	21	18
	D	5,210	448	3,068	557	498	403	131	66	19	20
1942	*A*	5,863	...	290	2,599	1,083	1,006	598	224	35	28
	B	8,769	...	1,636	3,780	1,674	858	444	244	74	59
	C	9,058	39	4,875	2,898	703	212	232	93	3	3
	D	8,074	336	5,344	989	688	421	173	73	31	19
1943	*A*	7,067	...	116	3,282	2,069	688	668	102	114	28
	B	8,247	...	380	4,048	1,844	831	689	287	97	71
	C	7,316	...	1,892	3,353	1,277	429	291	59	7	8
	D	6,647	45	1,716	3,194	1,013	173	320	146	29	11
1944	*A*	7,107	...	1	1,207	3,622	1,221	680	158	202	16
	B	6,029	...	35	1,396	2,586	1,385	448	64	81	34
	C	6,792	...	309	2,362	3,055	730	224	78	34	...
	D	3,020	54	193	681	1,174	458	312	80	48	20
1945	*A*	4,687	...	33	499	1,023	1,851	762	351	138	30
	B	5,640	...	1,555	374	1,862	1,097	469	193	12	78
	C	7,293	34	3,277	668	1,502	1,133	462	162	34	21
	D	3,769	67	1,783	138	589	792	248	69	58	25
1946	*A*	4,768	...	45	1,940	487	1,238	804	140	113	1
	B	6,630	...	735	3,040	333	1,059	928	446	80	9
	C	4,403	45	1,306	1,723	310	535	268	215	...	1
	D	4,024	52	1,334	1,270	471	582	246	69		.
1947	*A*	4,382	...	58	1,253	1,628	394	507	293	149	100
	B	3,589	...	1,052	1,015	600	264	337	175	94	52
	C	8,122	1	4,904	1,831	782	169	253	115	41	26
	D	3,721	73	1,969	658	442	171	159	136	71	42
Avg. all years	*A*	4,666	...	215	1,521	1,206	842	505	209	103	65
	B	5,806	...	1,069	2,015	1,223	721	436	225	74	43
	C	6,255	59	2,799	1,889	805	359	206	92	35	11
	D	4,093	403	1,807	858	448	307	147	72	33	18

olds, 1,224 five-year-olds, 432 six-year-olds, 208 seven-year-olds, 122 eight-year-olds, and 26 nine-year-old and older fish, or a total of 5,200; then the survivors from this group of year classes, after an interval of one year, would be the number of 5- to 9-year-old and older fish at the beginning of 1945, or 1,793 five-year-olds, 604 six-year-olds, 270 seven-

TABLE 4
SIZE OF HADDOCK POPULATION AT BEGINNING OF EACH YEAR[1]

Year	Age in years						
	4	5	6	7	8	9 and older	Total
1932	304	385	327	218	122	108	1,464
1933	2,276	235	286	260	101	120	3,278
1934	993	695	272	154	71	51	2,236
1935	832	602	616	104	67	39	2,260
1936	1,326	561	321	239	75	50	2,572
1937	1,064	634	242	136	86	24	2,186
1938	737	326	315	139	52	43	1,612
1939	884	354	180	114	63	46	1,641
1940	1,650	394	174	98	44	26	2,386
1941	1,544	932	267	176	43	58	3,020
1942	1,097	780	456	181	64	42	2,620
1943	1,950	728	498	198	94	39	3,507
1944	3,188	1,224	432	208	122	26	5,200
1945	1,482	1,793	604	270	77	52	4,278
1946	406	1,097	914	324	106	37	2,884
1947	1,305	360	490	246	132	38	2,571
Total	21,038	11,100	6,394	3,065	1,319	799	43,715
Average	1,315	694	400	192	82	50	2,733

[1] Values are the average of the number of fish of the particular age from Seasons *A* and *B* of the year in question, and of the number of fish of 1 year younger from Seasons *C* and *D* of the preceding year.

year-olds, 77 eight-year-olds, and 52 nine-year-old and older fish, or a total of 4,278.

Having already obtained the total number of 4-year-old and older fish at the beginning of the year (from Table 4), and having now computed the total number of 4-year-old and older fish at the end of each year (the number of 5-year-old and older at the beginning of the next year), all such totals were entered in Table 5.

Computation of the yearly diminution of the stocks being measured was then only a matter of subtracting the value representing the stock at the end of the year from the value representing the stock at the beginning of each year.

THE FISHERY REMOVALS OR "*C*"

A measure of the yearly decreases in population size of completely available fish from year to year thus had been obtained. Inasmuch as

TABLE 5

RELATIVE SIZE OF POPULATION OF CERTAIN AGES OF HADDOCK AT THE BEGINNING AND END OF EACH 15 YEARS

Year	Number of 4- to 9-year-olds at beginning of year	Number of 4- to 9-year-olds at end of year[1]	Decrease
1932	1,464	1,002	462
1933	3,278	1,243	2,035
1934	2,236	1,428	808
1935	2,260	1,246	1,014
1936	2,572	1,122	1,450
1937	2,186	875	1,311
1938	1,612	757	855
1939	1,641	736	905
1940	2,386	1,476	910
1941	3,020	1,525	1,495
1942	2,620	1,557	1,063
1943	3,507	2,012	1,495
1944	5,200	2,796	2,404
1945	4,278	2,478	1,800
1946	2,884	1,266	1,618
Total	41,144	21,519	19,625
Average	2,743	1,435	1,308

[1] End of year = number of 5- to 9-year-olds at beginning of following year.

the purpose of this study was to determine to what extent such decreases were associated with, or were the result of, the removals by the fishery, it was necessary next to determine how many fish the fishery had taken from the population in the various years.

The fishery removals for the years 1931–47[6] were first tabulated in terms of pounds of fish. Having also the average weights of these fish that were landed, the total numbers caught were easily computed. The total pounds and numbers are shown in Table 6, and the total numbers in Figure 6.

The numbers caught were then reduced to numbers of each age by utilizing the percentage-age compositions referred to earlier. After summarizing by size groups and season, the number of fish of each age removed by the fishery in each of the 17 years is shown in Table 7.

[6] The landings for the ports of Boston, Gloucester, New Bedford, Mass., and Portland, Me.

TABLE 6
TOTAL CATCH OF HADDOCK FROM NEW ENGLAND BANKS

Year	Millions of pounds	Millions of fish
1931	101.801	34.979
1932	86.706	32.348
1933	70.272	26.623
1934	39.683	15.617
1935	68.579	28.565
1936	73.496	31.489
1937	83.973	32.528
1938	80.202	33.570
1939	91.181	38.911
1940	81.676	31.345
1941	111.611	46.944
1942	97.786	41.299
1943	80.215	33.036
1944	84.265	29.062
1945	65.284	22.091
1946	90.802	32.678
1947	98.082	38.931
Total	1,405.614	550.016
Average	82.683	32.354

DECLINE IN THE SIZE OF STOCK AS ASSOCIATED WITH VARIATIONS IN THE CATCH.

In an earlier section of this paper the yearly declines in the relative size of the stocks of those ages of haddock that were fully available to the fishery were computed (Table 5). In the section just completed, the catch of fish of each age in each year has been computed (summarized in Table 7). By summing the catches of fish of 4 to 9 years of age inclusive, in each year, the numbers of fish that were removed from the corresponding stock between the beginning and the end of the years involved were computed. Thus, in Table 8 data are presented which represent:

(1) the decrease in the relative size of the stocks of 4- to 9-year-old fish from the beginning to the end of each of the 15 years (1932–46) in thousands of fish per day, and

(2) the number of fish removed by the fishery from these stocks during each of these 15 yearly intervals, in millions.

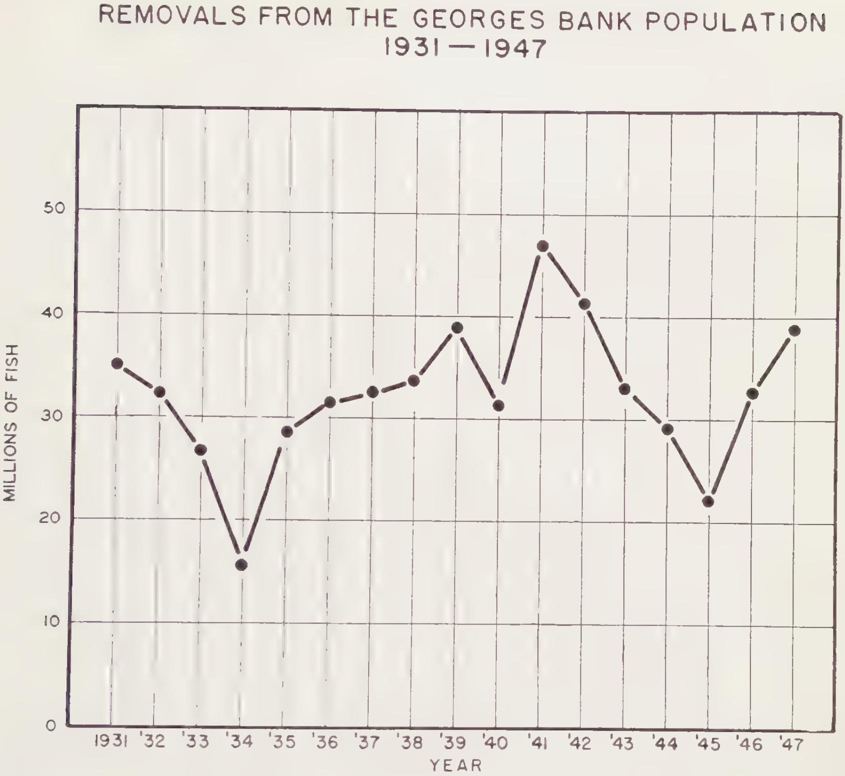


FIGURE 6.

CATCH OF GEORGES BANK HADDOCK IN TERMS OF NUMBERS OF FISH.

A casual observation of this table shows, in general, that in years during which large numbers of fish were taken from the Georges Bank population, there were also large declines in the population size from the beginning to the end of the year, and that in years during which small numbers were removed, the population changed but little.

These data have been plotted in Figure 7, with the "removal" or "catch (*C*)" as the independent variable, and with the decline, the change in population size from the beginning to the end of the year, as the dependent variable. This figure is plotted in a rather unusual manner, with values of the dependent variable being plotted below, rather than above the origin. This has been done inasmuch as values of the dependent variable (change in population size) are actually decreases rather than increases, and it has been found that this method of plotting is more easily interpreted by some people.

TABLE 7
AGE COMPOSITION OF CATCH, BY YEARS, IN MILLIONS OF FISH

Year	Age in years									Total
	1	2	3	4	5	6	7	8	9 and older	
1931	1.661	8.167	1.802	5.089	7.975	5.291	2.949	1.555	.490	34.979
1932	.099	1.712	21.139	2.008	3.051	2.383	1.079	.553	.324	32.348
1933	.210	7.366	5.218	8.648	1.791	1.346	1.030	.464	.550	26.623
1934	.296	3.807	4.470	2.889	2.518	.825	.482	.197	.133	15.617
1935	1.144	11.096	7.803	3.138	2.467	2.053	.415	.360	.089	28.565
1936	.828	11.449	10.171	4.629	1.803	1.153	1.140	.217	.099	31.489
1937	1.193	10.129	9.715	4.890	3.574	1.608	.815	.416	.188	32.358
1938	.961	18.453	6.866	3.304	1.568	1.312	.765	.198	.143	33.570
1939	.565	12.806	17.379	4.383	1.807	.804	.695	.272	.200	38.911
1940	1.895	6.692	11.061	7.261	2.188	1.286	.653	.191	.117	31.345
1941	.697	21.404	9.026	7.389	5.303	1.632	.860	.280	.353	46.494
1942	.290	13.106	14.877	5.938	3.648	2.131	.941	.205	.162	41.299
1943	.016	3.653	15.659	7.423	2.742	2.385	.688	.313	.157	33.036
1944	.054	.675	7.410	13.047	4.945	1.991	.435	.412	.093	29.062
1945	.101	7.046	1.698	5.285	4.862	1.941	.766	.223	.169	22.091
1946	.191	6.709	13.251	2.406	5.000	3.289	1.589	.224	.019	32.678
1947	.088	15.547	9.604	6.660	1.979	2.542	1.397	.690	.424	38.931
Total	10.289	159.817	167.149	94.388	57.221	33.972	16.700	6.770	3.710	550.016
Average	.605	9.402	9.833	5.552	3.366	1.998	.982	.398	.218	32.354

The straight line in Figure 7 was fitted to the data by the method of least squares and has the equation,

$Y = -.022 + .1135X$ where

X = millions of haddock of ages 4–9 years removed from the stock in each of 15 years by the fishery.

Y = decrease in relative population size of 4- to 9-year-old haddock during each of these 15 years in thousands of fish per day.

The coefficient of correlation measuring the degree of association between these two variables is 0.81. With 13 degrees of freedom this values proves to be highly significant (1 per cent level = 0.64). R^2 is about 0.66. Thus, it seems valid to conclude (under the assumption that the straight line best fits these data) that about 66 per cent of the variability in yearly decreases in population size, from the beginning to the end of the individual years, is explainable by the variations in the numbers of fish actually removed from the stock by the fishery. This value of 66 per cent is possibly a minimum estimate of the effect of the fishery, inasmuch as the index of abundance is probably not perfectly correlated with actual abundance.

No attempt was made in this treatment to determine whether some curved line fitted these data better than this straight line.

TABLE 8

DECREASE IN SIZE OF STOCK OF 4- TO 9-YEAR-OLD HADDOCK FROM THE BEGINNING TO THE END OF YEARS 1932–46, AND THE TOTAL CATCH OF THESE AGES IN EACH YEAR

Year	Decrease in stock thousands of fish per day	Catch in millions of fish
1932	.462	9.398
1933	2.035	13.829
1934	.808	7.044
1935	1.014	8.522
1936	1.450	9.041
1937	1.311	11.491
1938	.855	7.290
1939	.905	8.161
1940	.910	11.697
1941	1.495	15.817
1942	1.063	13.026
1943	1.495	13.708
1944	2.404	20.923
1945	1.800	13.246
1946	1.618	12.527
Total	19.625	175.720
Average	1.308	11.715

This line was arbitrarily extrapolated beyond the limits of the data towards the origin, although admittedly the exact position of the line where the removals are very small is unknown. It can be seen from Figure 7 that the intercept is practically at the 0.0 point. The position of this intercept, assuming that the population index actually represents the size of the population, poses interesting theoretical possibilities.

First of all, the suggestion is raised that within the ranges of population size and fishing removals represented by these data, the losses due to factors other than the fishing removals, i.e., to natural mortality, may be negligible. Such a possibility could theoretically be true under the conditions of an intensive fishery, where fishing removals would take many fish which would otherwise be removed by natural causes. When one takes into consideration the relative lack of bottom dwelling predators on Georges Bank that are large enough to consume haddock of 2–10 pounds and the apparent lack of any disease epidemic or serious parasitism in haddock over the 17-year period, this possibility does not appear quite so improbable.

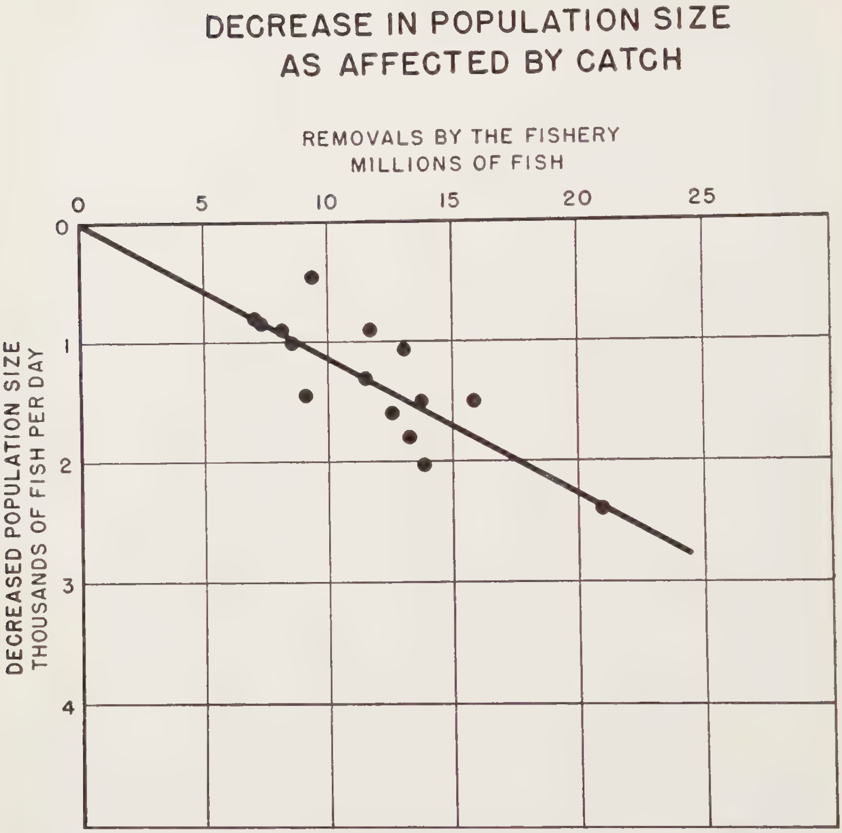


FIGURE 7.

THE RELATIONSHIP BETWEEN THE YEARLY REMOVALS IN MILLIONS OF HADDOCK AND THE DECREASE IN RELATIVE POPULATION SIZE FROM BEGINNING TO THE END OF THESE SAME YEARLY PERIODS IN TERMS OF THOUSANDS OF FISH PER DAY.

Secondly, the exact position of the line with populations of this general size, if the removals were greatly reducéd and even became zero, is unknown. If the extrapolation, as in Figure 7, happens to be an accurate representation of this relationship, then one would conclude that with no fishing removals, as would occur if fishing were to suddenly cease, there would be no decrease in the stock and thus no natural mortality. Theoretically, however, if fishing were to be considerably reduced suddenly, natural mortality would probably be greater than at present because some of the fish now being caught would be vulnerable to whatever causes of natural mortality are in operation. With populations of present levels but with very small fishing removals, the line would possibly curve toward the Y axis and intersect it at some point greater than 0.

The data and the ideas expressed here refer only to a heavily fished population and not to the relatively unfished populations of early days, or to the populations which would result if the sudden cessation of fishing were to continue for several years. In such populations, natural mortality would probably be greater yet, for such reasons as poorer nutrition of the larger stock, greater average age resulting in more deaths from senility, and so on.

This general situation is to be studied by various lines of approach in future studies. From the present study, however, we may conclude that the number of haddock caught in various years by the New England fleet markedly affected the subsequent population of haddock of corresponding ages on Georges Bank. Although it is generally assumed in many fisheries that the fishery does affect the stock, instances where such an effect has been demonstrated clearly are extremely rare. This analysis, in addition to demonstrating this relationship, is also of considerable value in providing the basic data that can be used in determining many other very important facts necessary for a broad understanding of the biometrics of the valuable New England haddock resource. Such facts include the actual number of fish present on the bank, fishing and natural mortality rates, growth rates, indices of the recruitment of young, the effect of various factors upon recruitment, and predictions as to the future abundance of this species. Investigations of these relationships are now being undertaken and will be reported upon soon.

ACKNOWLEDGMENTS

The system of collection of the basic biological and statistical data was developed by William C. Herrington, who until 1947 directed the North Atlantic Haddock Investigation. Employees of the United States Fish and Wildlife Service who aided the writer in summarizing the extensive data upon which this analysis is based include: James J. Miggins, Edgar L. Arnold, Jr., Louis D. Stringer, Frank A. Dreyer, Dorothy B. Monahan, Edward S. Phillips, Sterling L. Cogswell, Elizabeth V. Nugent and Margaret A. Reeves. Scales were read by Edgar L. Arnold, Jr., and the writer. Interpretation of certain data collected in early years was facilitated through the assistance of John R. Webster. The manuscript was reviewed critically by Dr. William F. Royce and Dr. Ralph P. Silliman of the Branch of Fishery Biology, United States Fish and Wildlife Service.

ONE DEGREE OF FREEDOM FOR NON-ADDITIVITY*†

JOHN W. TUKEY

Princeton University

INTRODUCTION

IN DISCUSSING the possible shortcomings of the analysis of variance, much attention has been paid to non-constancy and non-normality of the "error" contribution. (The recent papers in *Biometrics* by Eisenhart [4], Cochran [3] and Bartlett [1] discuss these matters and give references.) The present writer is usually much more concerned with and worried about non-additivity, and until recently has suffered from the lack of a systematic way to seek it out, and then to measure it. (Conversations with Frederick F. Stephan have contributed greatly to this development and presentation.)

The purpose of the present paper is to indicate such a way, when the data is in the form of a row-by-column table. (The professional practitioner of the analysis of variance will have no difficulty in extending the process to more complex designs.) We shall show how to isolate one degree of freedom from the "residue", "error", "interaction" or "discrepance", call it what you will. There are two known situations to which this single degree of freedom is expected to react by swelling:

(1) when one or more observations are unusually discrepant;
(2) when the analysis has been conducted in terms where the effects of rows and columns are not additive.

The first situation is quite familiar and requires little explanation. The second occurs often enough, but may not be noticed. An example may help to fix the ideas.

Let us construct an artificial example with 3 rows and 4 columns, with each entry contributed to overall, by rows, by columns, and by cells. Suppose that these contributions are as follows:

*Prepared in connection with research sponsored by the Office of Naval Research.

†Presented to the Biometrics Section and the Biometric Society at Cleveland, December 29, 1948.

in general				by rows				by columns				by cells			
1	1	1	1	4	4	4	4	6	1	−4	0	1	−2	1	0
1	1	1	1	−3	−3	−3	−3	6	1	−4	0	0	−1	2	−3
1	1	1	1	−3	−3	−3	−3	6	1	−4	0	0	−2	−1	0

Then the tables and corresponding analyses for the sum of all contributions are:

TABLE 1

ILLUSTRATIVE EXAMPLE IN ORIGINAL TERMS

Values and Means

	12	4	2	5	23	5.8
	4	−2	−4	−5	−7	−1.8
	4	−3	−7	−2	−8	−2.0
Sums	20	−1	−9	−2	8	
Means	6.7	−0.3	−3.0	−0.7		0.7

Analysis of Variance

	DF	SS	MS
Rows	2	140	70
Columns	3	157	52
R × C	6	26	4

Now let us square the entries and divide by 10, rounding to integers. The resulting tables and analyses are:

TABLE 2

ILLUSTRATIVE EXAMPLE IN TERMS OF SQUARES

Values and Means

	14	2	1	2	19	4.8
	2	0	2	2	6	1.5
	2	1	5	0	8	2.0
Sums	18	3	8	4	33	
Means	6.0	1.0	2.7	1.3		2.8

Analysis of Variance

	DF	SS	MS
Rows	2	24.5	12.2
Columns	3	46.9	15.6
R × C	6	84.8	14.1

Notice that all semblance of row or column effects have now vanished, although Table 1 showed large and significant effects. The use of the squared scale has concealed the real effects. (It may be argued that squaring numbers which range from plus to minus is unrealistic. The answer is that this *is* an extreme example, but one that can be slowly and smoothly changed into a very mild one. There probably is a difference in degree between this example and what happens in practice, but there is no difference in kind.)

PROCEDURE

How then do we isolate the single degree of freedom? The process is simple, and runs as follows:

(A) To the row-by-column table, already bordered with sums and means, add a new border of deviations of means from the grand mean (decimal places may be reduced, but the sums of deviations, by rows and by columns *must* be forced to vanish).

(B) Add an extra column (or row) and enter in each cell the sum of products of the deviations by columns and the entries in its row (or column).

(C) Accumulate the sum of products between the deviations of row (or column) means and the new entries of (B).

(D) Calculate the sum of squares of deviations by columns and by rows.

(E) Divide the square of the number from (C) by the product of the numbers from (D). This is the mean square (and also the sum of squares) for the single degree of freedom.

The process is illustrated on the same example below:

TABLE 3

SAMPLE CALCULATION

					Sums	Means	Deviations	Sums of x-products
	14	2	1	2	19	4.75	2.0	38.4
	2	0	2	2	6	1.50	−1.2	3.6
	2	1	5	0	8	2.00	−0.8	4.6
Sums	18	3	8	4	33		0.0	68.8
Means	6.00	1.00	2.67	1.33		2.75	6.08	
Deviations	3.2	−1.8	0.0	−1.4	0.0	15.44	50.9	

$$\text{(B)}: 14(3.2) + 2(-1.8) + 1(0.0) + 2(-1.4) = 38.4$$

$$2(3.2) + 0(-1.8) + 2(0.0) + 2(-1.4) = 3.6$$

$$2(3.2) + 1(-1.8) + 5(0.0) + 0(-1.4) = 4.6$$

$$\text{(C)}: 38.4(2.0) + 3.6(-1.2) + 4.6(-0.8) = 68.8$$

$$\text{(D)}: (3.2)^2 + (-1.8)^2 + (-0.0)^2 + (1.4)^2 = 15.44$$

$$(2.0)^2 + (-1.2)^2 + (-0.8)^2 = 6.08$$

$$\frac{(68.8)^2}{(15.44)(6.08)} = 50.9.$$

Assigning the mean square 50.9 to the degree of freedom for non-additivity, which is subtracted from "R × C", the analysis of variance of Table 2 becomes:

Rows	2	24.5	12.2
Columns	3	46.9	15.6
Non-additivity	1	50.9	50.9
Balance	5	33.9	6.8

Thus the obvious thing about the illustrative example was its non-additivity. The corresponding *F* value of 7.3 on 1 and 5 degrees of freedom is significant at the 5% level.

EXPLANATION

We have explained what we are looking for—non-additivity—and how to look—last section—but we have not explained what we are really doing. This we shall now try to do. Those experienced with single degrees of freedom may have already recognized the computation as a short-cut method of eliminating the single degree of freedom labeled by

$$\begin{vmatrix} 6.40 & -3.60 & 0.00 & -2.80 \\ -3.84 & 2.16 & 0.00 & 1.68 \\ -2.56 & 1.44 & 0.00 & 1.12 \end{vmatrix} = \begin{vmatrix} 2.0 \\ -1.2 \\ -0.8 \end{vmatrix} \cdot \begin{vmatrix} 3.2 & -1.8 & 0.0 & -1.4 \end{vmatrix}$$

where 6.40 = (2.0)(3.2), −3.60 = (−1.8)(2.0), 2.16 = (−1.8)(−1.2) and so on. We have used the products of the deviations of the row means and the deviations of the column means to label this single degree of freedom. Since the sum of each column and of each row is zero, this degree of freedom is orthogonal to rows and to columns. It must be a part of "R × C". This is what we did, but why?

Let us take a special case, where there are row contributions, and column contributions, *and nothing else.* We start with perfect additivity. If x_i is the column contribution (where i goes from 1 to c, the number of columns), and if y_j is the row contribution (where j goes from 1 to r, the number of rows), then the ij entry in the table is

$$a_{ij} = x_i + y_j .$$

Now let us start to analyze a slightly nonlinear function of the a_{ij} . Instead of a_{ij} , consider

$$f_\lambda(a_{ij}) = a_{ij} + \lambda(a_{ij} - a)^2$$

where λ is a small constant, and a is, for convenience, the average $\bar{x} + \bar{y}$ of all the a_{ij} . We find that we can write

$$f(a_{ij}) = [x_i + \lambda(x_i - \bar{x})^2] + [y_j + \lambda(y_j - \bar{y})^2] + \lambda(x_i - \bar{x})(y_j - \bar{y}).$$

The first two terms depend, respectively, on the column alone and on the row alone, so the last one contains all the *non-additive* effect due to analysis in terms of $f(a)$ instead of in terms of a. Notice that this non-additive effect is a multiple of

$$(x_i - \bar{x})(y_j - \bar{y}).$$

This means that it occurs in a single degree of freedom, which is identified in terms of $x_i - \bar{x}$ and $y_j - \bar{y}$.

We assumed no error of measurement, or the like, and we wrote $a_{ij} = x_i + y_j$ without an additional term. This means that the difference between the i-th column mean and the grand mean is

$$(x_i - \bar{x}) + \lambda\{(x_i - \bar{x})^2 - \overline{(x_i - \bar{x})^2}\}$$

which is nearly $x_i - \bar{x}$ when λ is small. Thus a satisfactory approximation to the single degree of freedom we want is that indicated by the coefficients

(column mean − grand mean)(row mean − grand mean).

This is exact for the combination of no error and a very slight change from a to $f(a)$, that is for no error and λ small. This fact plus empirical tests seems enough to warrant recommending general use of this single degree of freedom as a test of non-additivity.

WHAT OF SIGNIFICANCE?

Suppose that the test shows statistically significant evidence of non-linearity—what then? The simplest and laziest thing to do would be to forget the degree of freedom for non-additivity and go on and use the mean square for the balance in considering for example, the significance of the row effects. *This is not recommended*, for the following reasons:

(1) In general, results expressed in terms in which effects are additive apply in a broader region and are practically more useful.

(2) If the "error" or fluctuating contribution is not normally distributed, then it is not known whether or not the use of the balance mean square unduly inflates the apparent significance of other mean squares (for the case of a normally distributed fluctuating contribution there is no distortion of significance.)

For these reasons, the occurrence of a large non-additivity mean square should lead to consideration of a transformation followed by a new analysis of the transformed variable.

This consideration should include two steps:

(a) inquiry whether the non-additivity was due to analysis in the wrong form or to one or more unusually discrepant values;

(b) in case no unusually discrepant values are found or indicated, inquiry into how much of a transformation is needed to restore additivity.

The decision under (a) will depend on an examination of the data and all the background information available in the field—in particular the result of similar inspections of other experiments for non-additivity. What seems to be the best way of inspecting the results of a single experiment so far proposed is to plot the entries in the new column (of sums of cross-products) against the corresponding row means. A single unusually discrepant observation will tend to be reflected by one point high or low and the others distributed around a nearly horizontal regression line. An analysis in the wrong terms will tend to be reflected by a slanting regression line.

The figure shows such a plot, including 2s limits, for

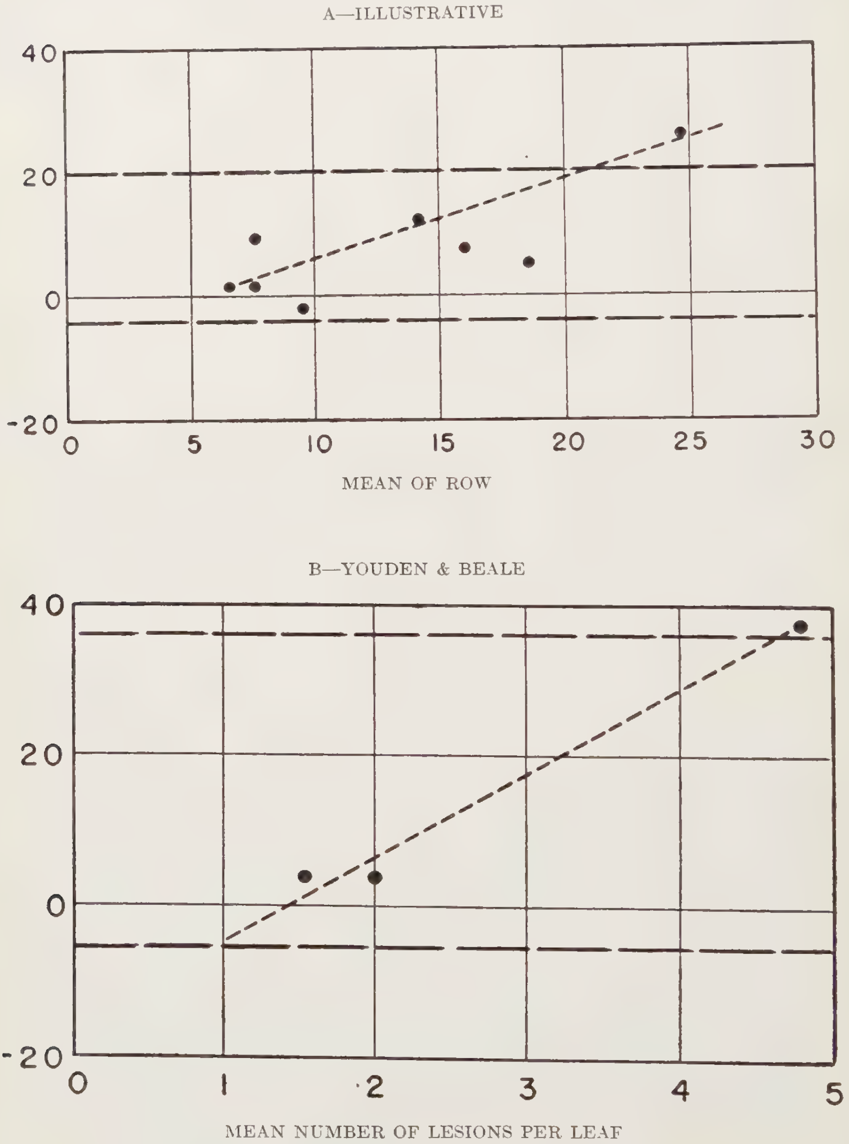
(A) the illustrative example worked above,

(B) Youden and Beale's data [6] as simplified by Snedecor [5, p. 44],

(C) Beall's experiment VI [2] on insect infestation, with plots

GRAPHICAL ANALYSIS OF NONADDITIVITY

(Ordinates are Sums of Cross Products, Dashed Lines are 2S Limits)

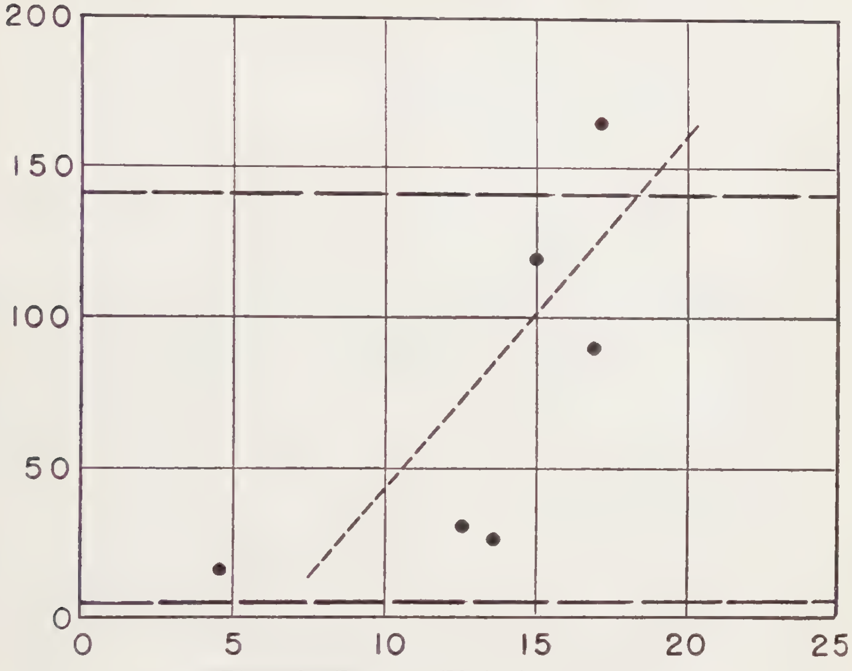

A—ILLUSTRATIVE

B—YOUDEN & BEALE

treated alike combined (analyzed in terms of numbers of insects).

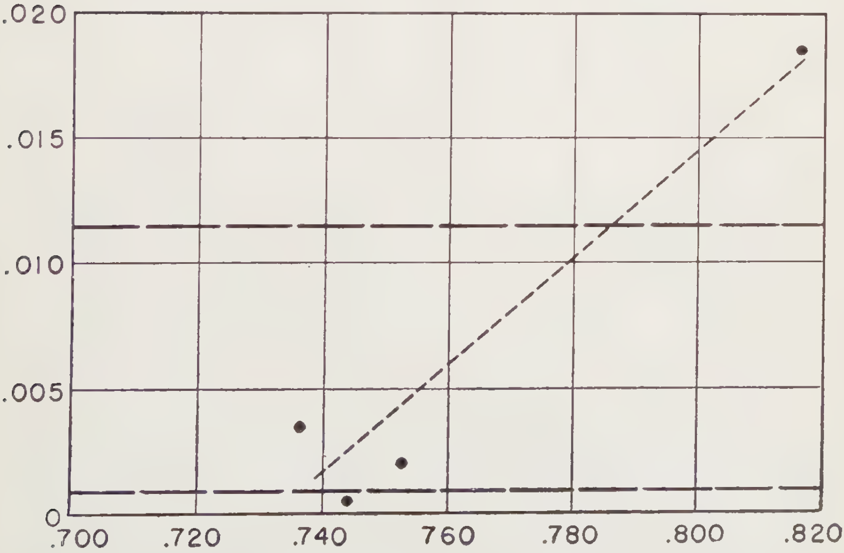
(D) Cochran's example [3] of an obviously discrepant value.

C—BEALL



MEAN NUMBER OF INSECTS PER PLOT-PAIR

D—COCHRAN



MEAN RATIO OF DRY TO WET GRAIN

The limits are set by the formula

$$\left(\begin{matrix}\text{average}\\ \text{cross product}\end{matrix}\right) \pm 2 \left(\begin{matrix}\text{sum of squares of}\\ \text{deviations of column means}\end{matrix}\right)^{\frac{1}{2}}\left(\begin{matrix}\text{mean square}\\ \text{for balance}\end{matrix}\right)^{\frac{1}{2}}$$

For the illustrative example (Case A), this becomes

$$15.53 \pm 2\ (15.44)^{\frac{1}{2}}(6.8)^{\frac{1}{2}} = 15.5 \pm 20.5 = -5.0 \text{ and } +36.0.$$

In every one of the four cases, the plotted points could be accounted for by non-additivity due to analysis in incorrect terms. Cases A and D can also be accounted for by a discrepant point. This suggests that it will be hard to make this distinction for single experiments on this scale. When several small experiments are available for analysis, agreement in signs of the slopes of the graphs or equivalently, the signs of the sums obtained in Step C may show up analysis in incorrect terms.

Why does the graph fail to decide about Cases A and D? The reason is simple—either explanation is plausible. If in Case A we alter the upper left-hand entry from 14 to 2, the analysis of variance becomes:

	DF	SS	MS
Rows	2	0.5	0.2
Columns	3	4.9	1.6
Non-additivity	1	0.2	0.2
Balance	5	12.6	2.5

Thus we see that our illustrative table of 3 × 4 entries *could* have perfectly well come from an additive situation where exactly one entry has been seriously disturbed.

Similarly in Case D, taken from Cochran's paper, if a nonlinear function is chosen so that

$$g(y) = \begin{cases} y, & .704 \leq y \leq .792, \\ .800, & y = 1.035, \end{cases}$$

then his table is converted into one where the F-ratio for non-additivity against balance is 0.8 instead of 27.6. We know that this table arose from an error in computation, but it *could* equally well have come from an additive table analyzed in the wrong terms.

In each case, the graphical solution has gone as far as it reasonably could in assigning responsibility for the non-additivity. While the graphical analysis is not certain to settle Step (a), it may be expected to be a big help.

AID IN CHOOSING A TRANSFORMATION

If it has been decided that the wrong terms had been used, then the actual size of the mean square for non-additivity must be useful for choosing an appropriate transformation. We lack experience with the more delicate use of such information, so that it seems appropriate to stop here with the following table which shows the connection between the *sign* of the final sum of products (which was +68.8 in the illustrative example) and the type of transformation which may then be appropriate.

TABLE 4

SIGN OF FINAL SUM OF PRODUCTS WHEN CERTAIN TRANSFORMATIONS ARE APPROPRIATE (VALUES OF x OR $x + a$ NON-NEGATIVE)

Transformed values which are additive*	Conditions needed	Sign when x is analyzed	Important special cases
x^p or $(x + a)^p$	$0 \leq p < 1$	$+$	$\sqrt{x}, \sqrt{x+1}$
	$p = 1$	0	(x)
	$1 < p$	$-$	x^2, x^3
$\log (x + a)$	(none)	$-$	$\log x, \log (1 + x)$

*Multiplication by a fixed constant and addition or subtraction of a fixed constant freely possible

While the removal of non-additivity by transformation usually tends to stabilize the variance, there may be cases where the variance is notably non-constant after transformation. In such cases, analysis of the transformed data using weights seems appropriate.

APPENDIX

VALIDITY OF THE ANALYSIS

This section is prepared for those who may feel that the method of obtaining the "single degree of freedom" may not produce quantities with the usual distribution.

The basic fact is this: If $u_1, u_2, \cdots, u_k; v_1, v_2, \cdots, v_m$ have some joint distribution, and if, for fixed $u_1, u_2, \cdots, u_k$, the distribution of $v_1, v_2, \cdots, v_m$ exists and is *always the same*, then the marginal distribution of $v_1, v_2, \cdots, v_m$ exists and, indeed, is the same, and, furthermore, $u_1, u_2, \cdots, u_k$ and $v_1, v_2, \cdots, v_m$ are independent. This can be established either by general considerations or by analytical detail.

To apply this in our case, let $u_1, u_2, \cdots, u_k$ be the row and column means, and let v_1 and v_2 be the sums of squares for non-additivity and for the balance. If the situation is additive, and the cell effects are normally distributed, and $u_1, u_2, \cdots, u_k$ are fixed, then v_1 and v_2 are independently distributed like σ^2 times chi-squares on 1 and $rc - r - c$ degrees of freedom. Hence v_1 and v_2 have these distributions, and are independent of all functions of row and column means. Thus the F-tests of rows, columns, or non-additivity against balance are valid.

In the presence of non-additivity and, or non-normality, the usual arguments indicate that the F-test is, if anything, conservative.

REFERENCES

[1] Bartlett, Maurice S. The Use of Transformations. *Biometrics* 3, 39-57, 1947.

[2] Beall, Geoffrey. The Transformation of Data from Entomological Field Experiments so that the Analysis of Variance becomes Applicable. *Biometrika* 32, 243-262, 1942.

[3] Cochran, W. G. Some Consequences when the Assumptions for the Analysis of Variance are not Satisfied. *Biometrics* 3, 22-38, 1947.

[4] Eisenhart, Churchill. The Assumptions underlying the Analysis of Variance. *Biometrics* 3, 1-21, 1947.

[5] Snedecor, George. *Statistical Methods*. The Collegiate Press, Ames, Iowa; 4th edition, 1947.

[6] Youden, W. J. and Beale, Helen Purdy. A Statistical Study of the Local Lesion Method for Estimating Tobacco Mosaic Virus. *Contributions from the Boyce Thompson Institute* 6, 437-454, 1934.

ON A STATISTICAL APPROXIMATION TO THE INFECTION INTERVAL

J. B. CHASSAN*

IN A PREVIOUS PAPER (2) the existence of strong correlation between the logarithms of the morbidity rates of a group of respiratory diseases for successive calendar month-pairs was demonstrated. The case rates involved pertain to the combined incidence of catarrhal bronchitis, acute coryza, acute catarrhal pharyngitis and laryngitis, and influenza, as diagnosed and reported in the United States Army. Where C_i is the case rate observed in the i-th calendar month, and C_{i+1}, the corresponding rate observed in the succeeding calendar month of the same year (or the same winter when i = December), the value of $r_{\log C_i \log C_{i+1}}$ for the twelve month-pairs averaged .84, each of the twelve coefficients being based upon some 38 observations, according to the number of years for which data were available for each month-pair. The purpose of the present paper is to relate some of the results obtained in connection with ref. (2) to the *law of mass action in epidemiology*, and to derive therefrom an estimate of the infection interval for an assumed period of immunity following infection, or conversely, an estimate of the period of immunity corresponding to a known infection interval. In connection with the actual numerical values presented, it should be noted that they pertain to a *group* of diseases and therefore can be interpreted only as average for the group as a whole.

The law of mass action in epidemiology states that the rate at which a contagious or epidemic disease spreads in a community is proportional to the product of the number of infectious individuals and the number of susceptibles in the community. If two consecutive time intervals are chosen such that the length of each interval is equal to the period between contact and case manifestation (i.e., the incubation period), a contact between an infectious person and a susceptible in the first interval will result in a new case during the second. Then the law of mass action may be written as

*The author wishes to acknowledge the helpful criticism and suggestions of Prof. John W. Tukey of Princeton University in connection with this paper.

$$C_{i+1} = \frac{1}{m} S_i I_i \tag{1}$$

in which

(a) C_{i+1} is the expected number of cases (or the case rate) during the $(i + 1)$-th period.
(b) S_i is the average number of susceptibles in the i-th period.
(c) I_i represents the average number of infectious individuals during the i-th period.
(d) m^{-1} is the proportionality constant reflecting such factors as the degree of crowding in a community, seasonality; more abstractly "infective power".

For the case in which the period of communicability following infection is relatively short, it is convenient to consider incidence in successive intervals whose lengths are each equivalent to the infection interval, rather than to the incubation period. The infection interval may be defined as the average period between the manifestations of two cases, one case resulting from contact with the other. It can be regarded as the sum of two components: first, the (average) time it takes for adequate contact to take place between a newly infected person and a susceptible, and then, the period between contact and manifestation. In such a case, we may replace I_i by C_i in equation (1), obtaining Soper's formula,

$$C_{i+1} = \frac{1}{m} S_i C_i \tag{2}$$

which gives the relationship between incidence rates in two consecutive periods whose lengths are each equal to that of the infection interval.

Soper (1) has also stated the relationship for the case in which the incidence rates are taken over successive periods of arbitrary length. If C_i is the case rate observed during the i-th month, and S_i is the average number of susceptibles in the i-th month, then

$$C_{i+1} = \left(\frac{S_i}{m}\right)^p C_i \tag{3}$$

where C_{i+1} is the incidence rate in the $(i + 1)$-th month, and p represents the number of infection intervals in one month.

If C_i is expressed as a daily incidence rate in terms of the number infected per day out of each 1000 population, and if n is the number of days of immunity following infection, then nC_i will give the average number per 1000 population who are not susceptible, by virtue of recent infection, during the month in which C_i is observed. On the assumption of general susceptibility in the population, the corresponding number of susceptibles per 1000 will then be given by

$$S_i = 1000 - nC_i \tag{4}$$

Substituting this value in (3), we obtain

$$C_{i+1} = \left(\frac{1}{m}\right)^p (1000 - nC_i)^p C_i \qquad \text{or}$$

$$\log C_{i+1} = p \log m^{-1} + p \log (1000 - nC_i) + \log C_i \tag{5}$$

Interpreting this equation in a statistical sense, i.e., as a regression function in which $x_{i+1} = \log C_{i+1}$ is regarded as the average value corresponding to a fixed observation of $x_i = \log C_i$, the data for the group of respiratory diseases under consideration indicates that the true regression curve of x_{i+1} on x_i increases monotonically with slowly declining slope over the actual range of observations. Apart from sampling differences a straight line of the form

$$x_{i+1} = a + bx_i \tag{6}$$

fitted by the method of least squares should lie close to the regression curve over the range of observed values of x_i, and the slope of the line, b, should very nearly equal the slope, β, of the secant which intersects the true regression curve at points corresponding to the lowest and highest of the observed values of x_i, respectively. An approximation to the infection interval can then be obtained by equating b, the slope of the linear regression of x_{i+1} on x_i, to β, the slope of the secant.

If C_{i_L} represents the lowest of the observed rates in the i-th month, and C_{i_u}, the highest; their substitution, in turn, for C_i in equation (5) above yield, as co-ordinates of the secant at the points of intersection with the mass action curve, the points,

$$\left(\log C_{i_L}, \log \left\{\left(\frac{1}{m}\right)^p (1000 - nC_{i_L})^p C_{i_L}\right\}\right)$$

and

$$\left(\log C_{i_u}, \log \left\{\left(\frac{1}{m}\right)^p (1000 - nC_{i_u})^p C_{i_u}\right\}\right)$$

respectively, where each ordinate is expressed as a function of the corresponding abscissa.

Then, from elementary analytic geometry the slope of the secant will be

$$\beta = 1 + p \left\{ \frac{\log \left(\dfrac{1000 - nC_{i_u}}{1000 - nC_{i_L}} \right)}{\log \, (C_{i_u}/C_{i_L})} \right\} \tag{7}$$

Solving for p, and substituting b for β,

$$p = (b \dot{-} 1) \left\{ \frac{\log \, (C_{i_u}/C_{i_L})}{\log \left(\dfrac{1000 - nC_{i_u}}{1000 - nC_{i_L}} \right)} \right\} \tag{8}$$

Upon applying formula (8) to the data of reference (2) for the twelve month-pairs, a median value of $p = 15$ was obtained on the assumption of three weeks of incidence as equivalent to the number of non-susceptibles, i.e., when $n = 21$. Since the incidence data were taken over monthly intervals, the corresponding estimate of the average infection interval is 2.0 days. On the assumption that $n = 28$, the median value of p is 11, and the infection interval, 2.8 days. Finally, if the assumption is made that $n = 42$, an average infection interval of 4.1 days is estimated. Thus in the neighborhood of the assumed values of n, the ratio of the period of immunity to the length of the infection interval is approximately 10 : 1.

Illustrating the procedure graphically, the chart given shows:

(a) a plotting of observed points corresponding to the February–March relationship

(b) a theoretical drawing of (5), represented by the curve MM', and interpreted as a regression curve

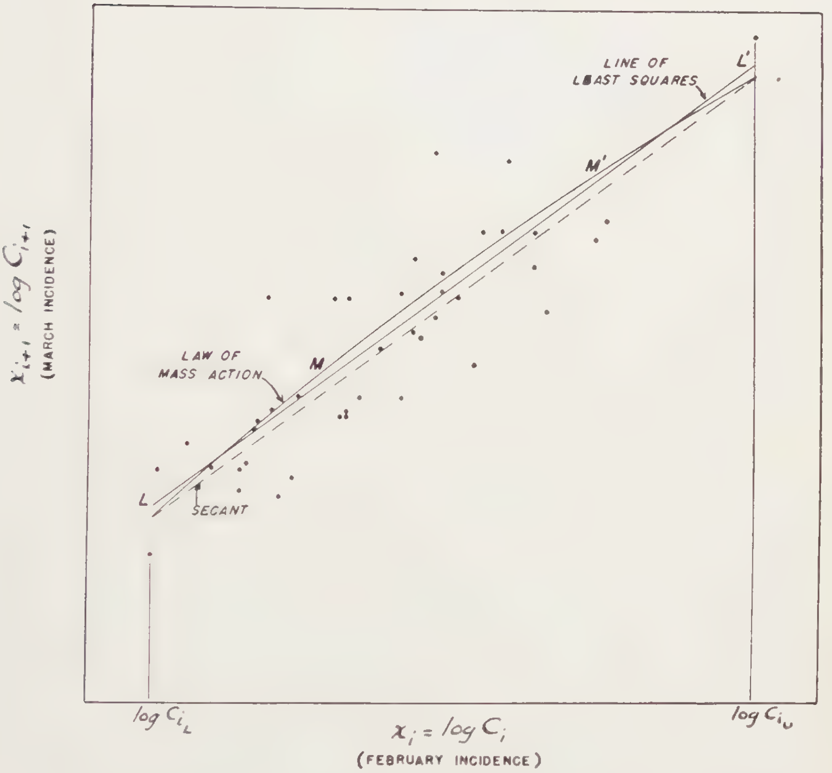
(c) the least squares linear regression of x_{i+1} on x_i, LL', fitted to the scatter of points

(d) a secant to the curve MM', drawn as a dashed line; the secant is drawn so that it intersects the curve to the left at the point whose abscissa is $\log C_{i_L}$, where C_{i_L} is the smallest of the observed values C_i, and to the right, at the point whose abscissa is $\log C_{i_u}$, where C_{i_u} is the largest of the observed values of C_i.

The position of the curve MM' in relation to its secant and to the least squares line LL', (again, apart from sampling errors) can be determined by formulating the vertical distance between MM' and the secant. By differentiation, both the maximum distance and the value of $\log C_i$ at which the maximum distance occurs can be determined. Thus if the equation of the secant is given by

$$\log C_{i+1} = \alpha + \beta \log C_i \,, \tag{9}$$

GRAPHICAL REPRESENTATION OF THE ESTIMATING RELATIONSHIPS IN THE APPROXIMATION TO THE INFECTION INTERVAL



THE SECANT TO THE LAW OF MASS ACTION CURVE, INTERSECTING THE CURVE AT EXTREMES OF x_i IS ASSUMED PARALLEL TO THE LEAST SQUARES LINE OF REGRESSION.

it will be found, by substituting $C_{i_L} = C_i$ in equation (9) and in (5) that

$$\alpha = p \log m^{-1} + p \log (1000 - nC_{i_L}) + (1 - \beta) \log C_{i_L} .$$

The distance from the secant to the curve will then be

$$\phi = p \log \left(\frac{1000 - nC_i}{1000 - nC'_{i_L}} \right) + (1 - \beta) \log \frac{C_i}{C_{i_L}} \qquad (10)$$

where b can be substituted for β, and p is obtained from (8).

The maximum value of ϕ can, of course, be obtained by differentiation with respect to C_i, or $\log C_i$ and equating to zero. Then the curve

MM' can be closely approximated from the position of the least squares line. Taking

$$M = L + \phi - 1/2 \max \phi$$

where, for a fixed value of log C_i, L is the corresponding value of log C_{i+1} on the least squares line, and ϕ is taken from (10), M is the corresponding value of log C_{i+1}, on the curve.

For an assumed value of p, equation (8) can be solved for n. From

$$\left(\frac{C_{i_u}}{C_{i_L}}\right)^{(b-1)/p} = \frac{1000 - nC_{i_u}}{1000 - nC_{i_L}}$$

we obtain

$$n = \left\{\frac{1000\left[1 - \left(\dfrac{C_{i_u}}{C_{i_L}}\right)^{(b-1)/p}\right]}{C_{i_u} - C_{i_L}\left(\dfrac{C_{i_u}}{C_{i_L}}\right)^{(b-1)/p}}\right\}$$

In applying the foregoing type of analysis the following modifications or limitations should be considered:

(i). We have assumed that for a fixed month-pair the infectivity factor, m^{-1}, is constant, except for random variation. Were it not for the fact of a declining number of susceptibles, S_i, with increasing C_i, as described by equation (4) above (i.e. if S_i were constant over the range of C_i), the mass action curve as given by equation (5) above would assume linear form with slope unity. But the declining value of S_i has the effect of causing the slope to drop with increasing C_i, so that apart from sampling errors, the slope of b (and of β) will be less than unity. This can be seen quite easily if we write equation (3) as

$$C_{i+1} = A_i C_i$$

Then, if A_i were constant for all C_i, a plotting of the curve (on a log-log scale) would yield a straight line parallel to

$$C_{i+1} = C_i$$

at a vertical distance of log A_i. But with the damping effect of the decline of susceptibles as C_i increases, A_i correspondingly decreases; and if, for example, log A_i is still positive the distance between the two lines decreases with increasing C_i, and it then follows that $\beta < 1$. The same result would, of course, apply when log A_i is negative.

Now if the situation were such that as C_i increases various preventive measures are taken which significantly reduce the infectivity factor, further damping will take place, and b will become smaller. To take this into account it would then be necessary to adjust upward the value of b, resulting in a corresponding increase in the length of the infection interval for each of the assumed values of n.

(ii). Equation (4) above implies that the entire population is potentially susceptible, and that the only immunes present at any given time, are those individuals who have gained immunity for a short period by virtue of recent infection. If, however, only a fraction, q, of the entire population are potentially susceptible, then instead of (4), it would be necessary to write

$$S_i = 1000q - nC_i \tag{11}$$

and substituting this, instead of (4) in (5), and in (7) and (8), it will be seen that for the same observed value of b, a somewhat longer infection interval would be estimated, depending on the degree of departure of q from unity.

(iii). Equations (4) and (11) will progressively lose accuracy as n gets very large. Thus, if the period of immunity were to last several months, these expressions would require modification to take account of variation in C_{i-1} , C_{i-2} $\cdots$.

References (3) and (4) listed below, and others listed in these references, discuss various aspects of the law of mass action of importance in connection with epidemic theory.

REFERENCES

1. Soper, H. E. The Interpretation of Periodicity in Disease Prevalence. *Jour. Roy. Statist. Soc.* 92, 34–61, 1929.
2. Chassan, J. B. The Autocorrelation Approach to the Analysis of the Incidence of Communicable Diseases. *Human Biology* 20, 2, 90–108, 1948.
3. Wilson, E. B. and Worcester, Jane. The Law of Mass Action in Epidemiology. *Proc. Nat'l Acad. Science* 31, 1, 24–34, 1945.
4. Wilson, E. B. and Burke, Mary H. The Epidemic Curve. *Proc. Nat'l Acad. Science* 28, 9, 361–367, 1942.

QUERIES

70 QUERY: I am carrying forward research on little known or on unknown tropical feedstuffs. For this research, rats, baby chicks and pigs are being employed. The unknown feedstuffs are evaluated singly and in combinations. I would appreciate your opinion on the proper method of statistical analysis for our data.

As an example and for brevity, here are some actual data from a pilot trial, together with the analysis of variance.

WEIGHT GAINS OF BABY CHICKS

	Treatment				
No. chicks	1	2	3	4	Entire sample
1	55	61	42	169	
2	49	112	97	137	
3	42	30	81	169	
4	21	89	95	85	
5	52	63	92	154	
	219	355	407	714	1695

ANALYSIS OF VARIANCE

Sources	D.F.	S.S.	M.S.
Lot means	3	26235	8745**
Individual	16	11559	722
Total	19	37794	

The F-test in the above case is highly significant indicating that we are not dealing with a single population. This method of analysis however does not provide us with a means of stating that treatment No. 3 is better than No. 1 or No. 2 is better than No. 4, etc. Could you provide us with the most valid method with which we could make these comparisons?

ANSWER: Happily this perennial question has been provided with an answer by Dr. John W. Tukey in the June issue of this Journal (Vol. 5: pages 99–114, 1949). Tukey's method indicates a gap between the first three treatments and the fourth. At a risk of less than one per hundred, one would reject the hypothesis of no difference between treatments No. 3 and No. 4.

There is not sufficient evidence to cut off the straggling mean of treatment 1 ($P = 0.17$). Finally, applying the F-test as indicated by Tukey, one does not reject the hypothesis that lots 1, 2, 3 are drawn from a common population ($P = 0.1$).

I assume that your experiment was conducted so that environmental differences were randomly distributed over all the chicks in the experiment; otherwise, there is no unambiguous answer to the question about the effects of treatments.

QUERY: 71 In an experiment in which one half of the controls reacted positively and one half negatively, it would seem that chi-square should be the same whether one uses the formula,

$$\chi^2 = 2(x - m)^2/m,$$

or the formula for the 2 × 2 table,

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)}$$

But this is not the case. Why?

For example, suppose 200 animals are divided equally among experimentals and controls. Then, according to the proposition under consideration, suppose 50 controls live and 50 die, and suppose 63 of the experimentals live and 37 die. Is the experimental procedure effective?

By the 2 × 2 table, $\chi^2 = 3.438$, not significant. But by the other formula, comparing the experimentals with a 1 : 1 ratio, $\chi^2 = 6.760$, highly significant. Why do not the two methods agree?

ANSWER: You have described two different experiments leading quite properly to different values of chi-square. In the first experiment there are only 100 animals, all treated experimentally. The assumption is made that in the untreated population the ratio of the numbers living and dying is 1 : 1. The hypothesis being tested is that the same ratio applies to the treated population; that is, that the treatment is without effect. The value of chi-square, 6.760, would lead to rejection of the hypothesis with P approximately 0.01. In this experiment there are no controls because the experimenter supplies the information about how controls behave.

The second experiment contains 200 animals, but half of them are used to get evidence about the behavior of the untreated population. Here the experimenter either has no knowledge of the behavior of the controls or is unwilling to rely on his knowledge. In this experiment, the hypothesis being tested is that the experimentals and controls have the same ratio, but the value of the ratio is not specified. The experimenter supplies less information than he did in the first experiment. The result is that the same number of experimentals, divided in the same ratio, lead to less certainty about the conclusion.

Querist feels that the chance division of the controls in the 1 : 1 ratio is equivalent to the 1 : 1 hypothesis which was set up in the first experiment. That this is not true may be clear if he considers the 95 percent confidence interval based on a sample of 100 equally divided in outcome. This interval is from 40 percent to 60 percent. The corresponding 99 percent interval is from 37 percent to 63 percent. Evidently the information supplied by such a sample of controls is far less than that furnished by the experimenter in postulating the 1 : 1 ratio for the population of controls.

72 **QUERY:** Hace un tiempo, se discutía en una reunión efectuada entre técnicos especialistas en maíz las exigencias para aprobar un híbrido o rechazarlo.—

Alguien sugirió aceptarlos cuando los rendimientos eran estadísticamente significativos.—

Y aquí comenzó la controversia. Otro técnico tomó la palabra para exponer su pensamiento al respecto. Dijo, que si se efectuaba un ensayo con todo cuidado, las exigencias para considerar un determinado híbrido estadísticamente superior a otro (altamente significativo), serían muy reducidos. Por ejemplo, un 3% de diferencias en los resultados, podría ser lo suficiente para que de acuerdo al análisis estadístico, se considere a un híbrido superior.—

Esto llevaría a un error, pues un 3%, en la práctica (en el gran cultivo) no tendría ninguna importancia, por lo que el procedimiento era erróneo. En cambio se mostró partidario de exigir un 10% de diferencia en los rendimientos y fijar un error standard de por ejemplo 6%.—

Desde luego, no sé a ciencia cierta quien tiene razón, por lo que recurro a Ud. a fin de que me evacúe la consulta. Puede hacerlo en inglés.—

ANSWER: Yield trials of various crops are usually conducted for one of two reasons: (1) to provide a test for a particular hypothesis or (2) to provide information which can be used as a guide in making recommendations over a range of soil and climatic conditions.

In the first instance an efficient experimental design and adequate replication are necessary so that the desired tests may be performed with the required precision. The number of replications and choice of design will, in part, be dictated by past experience as to soil variability, etc.

In the more general case where yield trials are conducted to provide information which will serve as a guide in making general recommendations, the situation is quite different. It is well established that different varieties respond differently in different years and at different locations. Therefore, varietal trials must be grown at several locations and in different years. Thus, there is little point in striving for "statistical significance" in each of the individual tests. An increase in number of replications for any single test will have little effect in reducing the magnitude of the variety × year or variety × location interaction.

The general practice in yield trials is not to select one or a few of the apparently superior items, but rather to discard a group of the poorer items. The items remaining are then tested further to provide additional information on performance.

If a number of varieties are tested over a series of years and locations, the outcome will almost certainly be a group of varieties which are so similar in yield and other characteristics that the differences among them will not be statistically significant. The best estimates of the relative value of the varieties in this group will be the actual averages obtained.

G. F. SPRAGUE

THE BIOMETRIC SOCIETY

By the time this number of *Biometrics* reaches you, each member of the Society will have received his free copy of our first Directory. Additional copies have been printed to send to new members as they are enrolled. It is available to non-members for 50 cents. Until a new edition is warranted, we propose issuing an annual supplement. As you will have discovered, the Directory includes a list of officers, the constitution of the Society, the Council by-laws, and the statutes of each region as well as the alphabetical membership list and a geographical summary. The information provided for each member includes his professional connection as recorded in the Secretary's office on June 15 and his major field of interest. Later, we hope to summarize the distribution of members among the different fields of interest. Although the Society has been in existence for less than two years, the geographical breakdown shows that we had 888 members in 33 different countries when the Directory went to press. The first and largest organized region was the Eastern North American, with 478 members. The other regions in order of formation were the British with 111 members, Western North American with 73 members, Australasian with 37 members, Indian with 43 members and French with 47 members. In addition, there were 99 members-at-large.

Since the last issue, the Council has approved the statutes of the Australasian, Indian and French Regions. These are already included in the Directory, so that they need not be reprinted here.

Developments in France are of unusual interest. The biometricians there have adopted a dual organizational plan in accord with a law of 1901 governing official French societies. They have formed the autonomous Société Française de Biométrie. At the same time they have formed the Region Française of the Biometric Society and provided that all full members of the Société Française de Biométrie shall be members of the Biometric Society. In view of this interesting development the tentative proposal of a joint French-Italian region has been abandoned. At the last meeting of the Société Française, on May 17 at the Laboratoire de Zoologie de la Faculte des Sciences, Paris, the following communications were presented: "La rehabilitation de l'homme moyen" by

M. Frechet, "Facteurs lateraux et facteurs sexuels dans la morphologie des empreintes digitales" by R. Turpin and M. P. Schutzenberger, and "Etudes biometriques sur le colibacille" by J. Dufrenoy.

Within the last months the following regional officers have been elected and confirmed by Council: British Region: Vice-President, J. W. Trevan; Secretary, D. J. Finney; Treasurer, K. Mather; Regional Committee, J. O. Irwin, J. I. M. Jones. Indian Region: Vice-President, P. C. Mahalanobis; Secretary, C. Radhakrishna Rao; Treasurer, Mohanlal Ganguli; Regional Committee, V. M. Dandekar, K. Kishen, K. R. Nair, U. S. Nair, V. G. Panse, P. B. Patnaik, B. Ramamurthy, R. V. Sukhatme, V. D. Thawani.

Since last November the Society has been provided with temporary headquarters in a pleasant room at 321 Congress Avenue in New Haven by the Department of Public Health of the Yale University Medical School. This room, however, will be required for new activities in the next academic year. Through the kindness of the Department of Applied Physiology, the Society has had the good fortune of obtaining a larger room at 52 Hillhouse Avenue in the main part of the University, and moved there on July 5. We would be very glad to welcome any visiting members at our new headquarters. We are very sorry to lose the services of Mrs. Elizabeth Weinman, who was Executive Assistant to the Secretary through June 30. The Society has benefited greatly from her efficient handling of the many details of the Secretary's office and wishes her well in her new undertaking. We have been fortunate in obtaining as her successor Mrs. Irving N. Fisher, who knows at first hand all of the countries where we have regions and most of the other countries where we have members.

NEWS AND NOTES

At the Raleigh branch of the Institute of Statistics there is a small news publication called the "Leaky Gasjet" which is printed irregularly depending upon the quantity of choice gossip acquired by its faithful seekers. The following excerpt was taken from the June, 1949, edition.

Dear Gasjet Editor:

I am a newly created Ph.D. in Experimental Statistics and I am worried because I expect to do consultation and I am afraid that the research workers will ask me questions that I won't be able to answer. What shall I do?

Phidler

Dear Phidler:

Here are a few simple devices which should prove useful to you in your consulting work. Relax, once you have mastered them you have absolutely nothing to worry about.

Research Worker: *Confidently.* I have done an experiment, Mr. Phidler, in which I have two plants, one of each variety, in each pot and fifteen pots. Can you tell me how to analyze it so as to show that Variety A is taller than Variety B? I realize *laughing selfconsciously* that this is a very elementary question but . . .

Phidler: *Frowning. Naturally as a new Ph.D. this is far too difficult a question for him, but he is not alarmed.* Just what do you mean by **taller?**

This illustrates both the Device of the Counterquestion and the Device of the Definition of Terms.

Research Worker: *A bit taken aback.* Taller? Well I mean bigger—not not bigger—

Phidler: *Sternly.* Come now, we cannot get anywhere unless we have specific, operational definitions.

Research Worker: Yes, of course. What I meant was I measured the height of each plant and—

Phidler: The external or the internal height? *He pauses, but Research Worker is unable to answer.* A similar problem came up in the Jour.-Roy-Stat-Soc-Supple-eleventy two-page 476.

The Device of the Non-Existent Reference

Research Worker: *Awed.* What was that reference again?

Phidler: No matter. It's by Gregory Hairshirt. I knew Hairshirt in kindergarten—an idiot—his papers were demolished by Smirkley Annals of Applied Human Genetics. Let's get back to our little problem.

The Device of Complete Familiarity with Everyone and Everything

Research Worker: *Relieved.* Yes, Yes. Now I thought this design—

Phidler: Design? *Laughs* Yes **design.** You realize of course that you should have used a cuboidal lattice in this experiment.

The Device of the Wrong Design

Research Worker: I—well I didn't know—

Phidler: *Aloud to the walls.* How do these research workers expect us to get anything out of their data when they use any old design. Ah well, I suppose we can work it out by matrix methods. Tell me, what is the Cost Function for height in this problem?

The Device of the Unnecessary Complication

Research Worker: Cost? I don't know—I thought this was a simple *sob*! problem—but after all I'm only a miserable research worker and not a statistician *alas*!

Phidler. *Benevolently.* Now, now, don't cry. I will help you. This is really a very simple problem.

The Device of Reversing your Field

Research Worker. *On his knees.* For you, perhaps, O Master. *The research worker is now in the proper frame of mind for consultation. From here on in Phidler can do ANYTHING.*

AFRICA—Among our new members, **Henri Marchand**, Dakar, Sénégal, West Africa, writes, "My researches are purely theoretical in the field of mathematical genetics. As soon as my present studies on the part that a single body can have on the evolution of a population are advanced, it will give me great pleasure to send you a report on the results at which I will have arrived."

AUSTRALIA—**Helen Turner** had plans all completed to attend the Second International Biometrics Conference in Geneva and to spend six months in Cambridge. Unfortunately family illness has intervened and the trip has been postponed. **D. B. Duncan** is busy developing a teaching program in Statistical Methods in the University of Sydney. Three new courses have been set up and the first graduate in Agricultural Science with Honors in Statistical Methods, **J. A. Morris,** took his degree this March and is now working in animal genetics in the Division of Animal Health and Production of C.S.I.R.O. **H. O. Lancaster** of the Commonwealth Health Department has just completed a year's study in England and is now on his way back to Australia. **E. A. Cornish** has a new F_1 to carry on the statistical tradition. **C. W. Emmens**, author of the recently published Principle of Biological Assay, is coping well with a large demand for presentation of papers to scientific societies in Sydney.

FINLAND—**Leo Törnqvist**, Chairman of the Institute of Statistics, University of Helsinki sends a brief note. He writes, "The Institute of Statistics in the University of Helsinki was founded in 1945, but has started its activity only in 1947. Its Chairman is the professor in statistics of the University, and an M.A. works there as assistant. The Institute is partly a statistical library, partly an advisory and direction office for the students of statistics. In addition the chairman, the assistant, and the more progressed students work with special statistical researches for outsiders. The received tasks have chiefly been from the branches of population—prognostics and analysis of economic time-series. The teaching in statistics belongs in the University under the Faculty of the Political Sciences. The student can choose statistics in the M.A.-examination for his chief subject or for one of his side subjects. After the M.A.-examination it is possible to go on with the studies as far as to the doctoral thesis. My special interests in statistics are the theoretical and economical problems."

INDIA—**D. N. Nanda** has taken up the position of a Statistician for Indian Army Ordnance Corps. He writes, "In this capacity I am to conduct Applied Research on the following subjects: (1) Design and Analysis of Experiments, (2) Quality Control, (3) Sampling Surveys (including inspection methods). There are a number of other topics on which I may have to work from time to time." He would appreciate being informed of the latest developments in these fields.

UNITED STATES—On February 1, **Alexander G. Ruthven**, president of the University of Michigan, announced the establishment of the Institute for Social Research. "The institute will be directed by **Rensis Likert** and will provide a unified administration for two units already existing at the University, the Survey Research Center and the Research Center for Group Dynamics. **Angus Campbell** will succeed Mr. Likert as Director of the Survey Research Center, which will continue its major programs of research in such fields as: studies of economic behavior and motivation; studies in human relations and organization; studies of the American public's understanding of major national and international issues; and the development of sampling survey methodology. **Dorwin Cartwright** will continue as Director of the Research Center for Group Dynamics. As a part of the Institute for Social Research this group will continue its program of research on the factors influencing productive and harmonious group functioning. It will continue its studies on human relations in industry, leadership, communication within groups, intergroup relations, and the social satisfaction of community life. As a result of the joining of the two centers, the Institute is better able to bring to bear quantitative and experimental research methods on complex and important social problems. Research findings of the Institute are communicated not only through teaching and scientific publications, but also through consultation and training in various organizations. The staff of the Institute includes over 350 persons engaged in full time or part time work. Approximately 125 of this number are located in Ann Arbor. Although most of the professional staff are social psychologists, various other social sciences are represented." **Melville A. Taff, Jr.**, formerly with the Louisiana State Department of Health, New Orleans, is now with the Territory of Hawaii Department of Health, Honolulu, as Chief of the Bureau of Health Statistics. Mr. Taff writes, "The Bureau is being expanded to provide statistical service for the entire department. Additional tabulating equipment has been ordered and more statistical personnel will be added as necessary. A central statistical service unit is the prime objective. An Act patterned after the Uniform Vital Statistics Act was passed at the 1949 session of the Legislature and now awaits the signature of the Governor. Once signed one of the first moves will be to consolidate small and sparsely populated registration districts and wherever possible and practical to appoint the local health officer as local registrar. Office methodologies are being reviewed and revised procedures are being written." **Paul T. Bruyere** formerly with the Army Institute of Pathology is now with the Division of Tuberculosis, United States Public Health Service. He, with **Martha**

Bruyere, is making a study of the early development of tuberculosis among student nurses. **Jack Chassan** also joined the United States Public Health Service and is working with the Bruyere's on the student nurse study. He recently left the Office of the Surgeon General, Department of the Army. **Allen B. Burdick** is now Assistant Professor of Agronomy, University of Arkansas, Fayetteville. He is initiating research in the development of grain and forage types of sorghum and will teach a course in the genetics of plant breeding. His theoretical research will continue to emphasize the mathematical aspects of quantitative inheritance. Mr. Burdick was with the Atomic Energy Commission at the Genetics Division, University of California, Berkeley. **H. M. C. Luykx** is resigning his position as Associate Professor of Preventive Medicine, at New York University College of Medicine, to accept appointment as Biometrician for the Atomic Bomb Casualty Commission in Japan. The Commission operates under the Committee on Atomic Casualties of the National Research Council, Washington, by directive of the President, and is sponsored by the Atomic Energy Commission. Mr. Luykx will be stationed in Japan for about two years, where he will make his home in Kure, with frequent visits to Hiroshima and Nagasaki. **R. L. Murphree** recently resigned his position with the Bureau of Dairy Industry at Jeanerette, Louisiana, to accept a position as Associate Professor of Animal Husbandry at the University of Tennessee. **Kenneth S. Cole,** formerly with the Institute of Radiology and Biophysics of the University of Chicago, is now Scientific Director of the Naval Medical Research Institute at Bethesda, Maryland. **Theodore A. Bancroft** has joined the staff of the Iowa State College Statistical Laboratory as Associate Professor—July 1, 1949. **Gobind Ram Seth,** on a four months' leave of absence from the Statistical Laboratory at Ames, flew early in July to visit the statistical institutions in Sweden and England, before returning to Delhi, India, where he will be teaching. **Oscar T. Kempthorne** was married in Vancouver, British Columbia, Canada, on June 10, 1949, to Miss Valda M. Scales of Coogee, New South Wales, Australia. Professor and Mrs. Kempthorne will be at home at 127 Stanton, Ames, Iowa, sometime in July.